Network generation and analysis for the investigation of host-pathogen interactions in tuberculosis

Holifidy Arisoa Rapanoël
RPNHOL001

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN
In fulfilment of the requirements for the degree

MSc in Bioinformatics

Faculty of Health Sciences
University of Cape Town

14 August 2012

Supervisor: Assoc. Prof. Nicola Mulder
Clinical Laboratory Sciences, University of Cape Town

Declaration

I, Holifidy Arisoa Rapanoël, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: .....

Date: .....

Abstract

The outcome of the infection by *Mycobacterium tuberculosis* (*Mtb*) depends greatly on how the host responds to the bacteria and how the bacteria manipulates the host, which is achieved through protein-protein interactions. This project aims to predict protein interactions between human and *Mtb*. The interologs method based on experimentally verified intra-species and inter-species interactions was used to predict human-*Mtb* protein interactions. They were further filtered using known host-pathogen interactions and genes that are differentially expressed during infection. This yielded a total of 190 interactions. Analysis of the subcellular location of the human-*Mtb* interactions showed that they are potentially feasible. The biological processes of the predicted human proteins suggested their involvement in apoptosis and production of nitric oxide, whereas those of the *Mtb* proteins were relevant to the intracellular environment of *Mtb* in the host. Mapping the proteins onto KEGG pathways highlighted proteins belonging to the tuberculosis pathway and also suggested that *Mtb* proteins might use the host to acquire nutrients. Finally, the predicted interacting *Mtb* proteins were enriched in previously predicted drug targets. Taken together, the predicted interactions could shed light on the interplay between *Mtb* and its human host.

Acknowledgements

I am sincerely grateful to Prof. Nicola Mulder for the supervision and guidance she showed throughout the time it took me to complete this project.

My thanks also to Dr. Gaston Mazandu for sharing his research with me. All my colleagues of the Computational Biology group who in one way or another contributed to the realisation of this project.

My thanks must go also to all the Malagasy in Cape Town. It really helps to feel like home.

Last but not least, my family for supporting me during this process, thank you so much.

Contents

Chapter 1

Introduction

Tuberculosis is an infectious disease that has plagued humans for thousands of years. Indeed, examination of ancient mummies from Egypt showed signs of tuberculosis [94]. In the 1700s and early 1800s, tuberculosis was the largest cause of death in Western Europe and the United States. It is still a deadly disease worldwide. Tuberculosis is one of the top three disease killers alongside HIV/AIDS and malaria. According to the World Health Organization (WHO), tuberculosis affects one-third of the world's population and 1.7 million people died from it in 2009 [143].

1.1 Tuberculosis transmission

Tuberulosis (TB) is an airborne disease caused by a bacterium called *Mycobacterium tuberculosis (Mtb)*. The bacilli enter the body via the respiratory tract by inhalation of contaminated droplets from people who have an active *Mtb* infection. In very young children and people with a weakened immune system, the organisms can grow and multiply immediately, causing primary tuberculosis. Up to 50% of the time the infection is cleared through a robust innate immune response and the host does not develop disease. The surviving bacilli are engulfed by alveolar macrophages and dendritic cells and can remain alive in those cells in a dormant state. In 10% of cases, the bacilli start to multiply and cause active disease. It generally happens to immunocompromised individuals such as HIV/AIDS infected people. In the remaining 90%, the bacilli remain inactive and will not cause any further problems [2].

Although residing mainly in the lungs causing pulmonary TB when active, the bacteria can spread to other parts of the body via the blood stream causing extra-pulmonary TB or

miliary TB. The kidneys and lymph nodes are the most common sites for extra-pulmonary TB. Tuberculosis can also affect the pleura, liver, spleen, bones and joints, heart, brain, genital organs, meninges, peritoneum and skin [121].

Tuberculosis transmits only from person to person. Persons with active tuberculosis contaminate the air when they expel bacteria by coughing, sneezing, or even speaking. The bacteria are able to survive in the air for several hours. Another person breathing them in may then become infected. However, a person who has latent TB does not expel bacteria into the air and therefore, cannot spread the disease. The most common symptom of tuberculosis is a cough, but patients may also have cold sweats, with or without fever, weight loss, chest pain and respiratory insufficiency.

1.2 Tuberculosis diagnostic and treatment

Diagnosis of tuberculosis includes chest x-ray, the tuberculin skin test, or examination and culture of sputum samples. The tuberculin skin test, also known as Mantoux or purified protein derivative (PPD) test can detect latent infection. However, BCG vaccination can also lead to a positive test, making PPD a less trusted method [113]. Microscopic detection of bacteria using sputum samples gives faster results but is less accurate than culture. Traditional cultures actually take many weeks to give results because of the slow growth of the bacteria. Molecular techniques detecting antimicrobial resistance markers in samples or culture are also available, giving results within 24 hours. Rapid and reliable diagnostic methods are expensive and therefore used only in developed countries. There is still a need to develop faster and less expensive tests to confirm TB cases.

Tuberculosis is a curable disease but the treatment takes time. Treatment of active tuberculosis consists of a combination of first-line drugs for two months followed by the combination of isoniazid and rifampicin for four months. First-line drugs include isoniazid, rifampicin, streptomycin, pyrazinamide and ethambutol [95]. The therapy must not be interrupted even if the patient feels well because the bacilli are not completely eliminated and failure to comply contributes to the development of drug resistance. In order to ensure that the treatment is fully completed, the WHO implemented a program called DOTS (directly observed treatment short-course) where patients are observed when they take their pills [96]. The treatment is extended when there is resistance to at least rifampicin and isoniazid, and the use of second-line drugs is required. This characterizes multi-drug resistant tuberculosis or MDR-TB. Extensively drug-resistant tuberculosis or XDR-TB is defined as MDR-TB and resistance to the fluoroquinolones and one of the three injectable second-line agents

(amikacin, kanamycin or capreomycin) [64]. MDR and XDR-TB need more complicated and costlier treatment. Three cases of totally drug-resistant tuberculosis (TDR-TB), a form that is resistant to all prescribed medications, have also been reported. The first case involved two patients in Italy in 2007 [89], the second involved fifteen patients in Iran in 2009 [137]. The latest case has been discovered in India in 2011 [132].

1.3 *Mycobacterium tuberculosis*

Discovered in 1882 by Robert Koch, a German physician and bacteriologist, *Mycobacterium tuberculosis* (*Mtb*) is an obligate human pathogen belonging to the genus *Mycobacterium* (*M.*). It is a rod-shaped, non-motile, non-sporing, aerobic and acid fast bacillus with a size varying from 0.3 to 0.6 μm in diameter and from 1 to 4 μm in length. The bacillus slowly replicates with a generation time of about 24 hours [42]. Although staining with difficulty with crystal violet, *Mtb* is classified as a Gram-positive bacteria. It has a unique cell wall composition consisting of mycolic acids linked to underlying arabinogalactan and peptidoglycan. This cell wall also contains different types of unique lipids such as lipoarabinomannan (LAM), lipomannan (LM) and phosphatidyl-*myo*-inositol mannosides (PIM). These lipids appear to play an important role in the virulence of *Mtb* [66]. Several lipoproteins and glycoproteins, such as lpqH (19 kDa), pstS1 (38 kDa), and lprG (Rv14llc) are also found in the cell wall, where they manipulate bactericidal mechanisms and play a role in the virulence of *Mtb* [149, 52].

1.3.1 Mycobacterial strains

The *Mtb* complex includes *M. tuberculosis*, *M. bovis*, causing tuberculosis in cattle, *M. bovis* BCG, *M. africanum*, *M. microti*, affecting voles [50], and *M. pinnipedii*, found in seals and sea lions [35]. Most of the time, human tuberculosis is caused by *Mtb*, although *M. bovis*, can also cause tuberculosis in human. *M. tuberculosis* consists of six main strain lineages associated with particular geographic regions: the Indo-Oceanic lineage is mainly found all around the Indian Ocean; the East-Asian strain lineage occurs in East Asia, Russia and South Africa; the East-Africa-Indian strain lineage is mainly found in the Indian subcontinent and in East Africa, the Euro-American strain lineage, in Europe and the Americas; and the West-African strain lineages, consisting of two lineages, occur almost exclusively in West Africa [51]. *Mtb* strains vary in phenotype, virulence and immunogenicity [51].

The laboratory strains H37Rv, H37Ra, Erdman and the clinical strain CDC1551 are the most

Function	Number of genes
Virulence, detoxification, adaptation	99
Lipid metabolism	233
Information pathways	229
Cell-wall and cell processes	708
Stable RNAs	50
Insertion sequences and phages	149
PE and PPE proteins	170
Intermediary metabolism and respiration	894
Proteins of unknown function	272
Regulatory proteins	189
Conserved hypothetical proteins	1051

Table 1.1: Functional classification of H37Rv genes

widely studied *Mtb* strains. The H37 strain was isolated from a human patient in 1905 and later dissociated into an avirulent strain (H37Ra) and a virulent strain (H37Rv) [73]. The strain CDC1551 or "Oshkosh" strain was isolated from a 21-year-old male clothing factory worker located in the Kentucky/Tennessee region (US). Valway *et al.* [134] showed that this strain was highly contagious, the index patient infected approximately 80% of his co-workers and social contacts. They also demonstrated that the CDC1551 strain was highly virulent in mice: after 20 days of infection, mice infected with the CDC1551 strain presented 100 times higher numbers of bacilli in their lungs than those infected with the Erdman strain.

1.3.2 The *Mtb* genome

The genome of the H37Rv strain was sequenced in 1998 by Cole *et al.* at the Sanger Centre (Hinxton, UK) [33]. The sequence initially contained $4,411,529$ base pairs (bp), with a G + C content of 65.6%. They identified 3,974 genes including 3,924 protein-coding genes and 50 genes encoding stable RNA. 40% of the predicted proteins were functionally annotated, some information or similarities were found for another 44% and 60% had no known functions. Upon re-annotation of the genome in 2002, 82 additional genes have been included, four sequencing errors have been corrected, changing the sequence length to $4,411,532$ bp [26], and 52% of the proteome (2,058 proteins) has been annotated. Table 1.1 shows the functional classification of the H37Rv genes after re-annotation.

Cole *et al.* revealed the existence of two new gene families that form about 10% of the

coding capacity of the genome [33]. These are the PE and PPE gene families that are found in mycobacteria only. They were so named because of the motifs Pro–Glu (PE) and Pro–Pro–Glu (PPE) found near the start of their encoded proteins [33]. These proteins are thought to play a role in evasion of host immune responses [33, 68].

In 2002, Fleischmann *et al.* at The Institute for Genomic Research (TIGR) sequenced the genome of the *Mtb* CDC1551 strain and compared it to that of the H37Rv strain [48]. They found that out of the 37 insertions contained in the H37Rv strain relative to strain CDC1551, 8 are complete open reading frames (ORFs). The CDC1551 strain contained 49 insertions relative to the H37Rv, 17 of them are complete ORFs. They also found that H37Rv contains 16 copies of IS*6110*, whereas CDC1551 had only 4 copies. IS*6110* is an IS3-type insertion sequence and principal epidemiological marker for *Mtb*. They identified regions differing in the copy number of several genes between the two strains. They also identified 1,075 single-nucleotide polymorphisms (SNPs) between the two genomes. The CDC1551 strain has a genome size of 4,403,837 bp, an average G + C content of 65.6% and 4,189 coding genes.

1.3.3 *Mtb* and the immune response

The immune response protects the host against infection, and comprises the innate immune system and the adaptive immune system. The innate immune response is the first line of defence of the body against pathogens, providing immediate defence against infection. Several receptors on the surface of the phagocytes (macrophages and dendritic cells) recognize *Mtb*. These receptors include the complement receptors, the mannose receptor (MR), dendritic cell-specific intercellular adhesion molecule-3 (ICAM-3)-grabbing non-integrin (DC-SIGN) (CD209), surfactant protein A (Sp-A) receptor, the class A scavenger receptor, the toll-like receptor (TLR) and mannose-binding lectin (MBL) [70]. Thus, the response depends on the specific receptor that helps a phagocyte ingest *Mtb*. If the receptor is a signalling pathogen recognition receptor (PRR) such as TLR or NOD-like receptor (NLR), the binding leads to the induction of an inflammatory response [130]. Binding to TLRs induce signalling cascades, most commonly NF-κB and MAPK pathways. Activation of the MAPK pathways occurs through a series of phosphorylation. MAP kinase kinase kinase (MKKK) first phosphorylates MAP kinase kinase (MKK), which in turn phosphorylates MAPK. MAPK then translocates to the nucleus where it activates by phosphorylation proteins required for inflammatory responses and apoptosis. For the NF-κB pathway, TLR stimulation also leads to phosphorylation events which activate the IκB kinase kinase (IKK) complex. IKK phosphorylates the α subunit of inhibitor of NF-κB (IκB), which is polyubiquitinated and degraded, allowing the nuclear translocation of NF-κB where it stimulates transcription of

genes required for pro-inflammatory responses, such as cytokines and chemokines, and genes involved in cell proliferation [130].

On the other hand, if the receptor is an endocytic PRR such as MR, the binding leads to phagocytosis [118], which is the process by which many cells ingest microbial pathogens and apoptotic or necrotic corpses [112]. Microbes are engulfed into a vacuole called the phagosome, which then undergoes a maturation process beginning with recruitment of activated Rab5, which is a GTPase. Activated Rab5 is necessary for maturation of the phagosome through the recruitment of a large number of effector proteins: the phosphatidyl 3-kinase (PI3K) hVPS34 that produces phosphatidylinositol-3-phosphate (PI3P) at the phagosome surface, which in turn recruits proteins containing PI3P motifs, including the early endosomal antigen 1 (EEA1), essential for fusion between early endocytic vesicles [4]. Rab5 is replaced by Rab7 in a process called Rab conversion, which mediates the conversion of the early endosome to late endosome. This is required for the fusion of the phagosome with the lysosome to form a phagolysosome. The phagolysosome is an acidic environment enriched in proteases, and conditions that promote bacterial degradation [4]. Meanwhile, the α and β chain of MHC (major histocompatibility complex) class II, along with the invariant chain (Ii or CD74) of the molecules are synthesized and assembled in the endoplasmic reticulum (ER), transported to the golgi apparatus, and from there to the phagolysosome. Once in the phagolysosome, the invariant chain is cleaved leaving a small peptide, the class-II associated invariant chain peptides or CLIP, bound to the class II molecule. HLADM, a vesicle membrane protein, catalyzes the removal of the CLIP peptide from the peptide binding groove of the MHC molecule and the binding of the pathogen-derived peptides. The MHC class II-Antigen complex is then transported to cell surface where it can be recognized by the antigen receptors of $CD4^+$T cells. Bacterial proteins processed in the cytosol by the proteasome are presented by MHC class I molecules to $CD8^+$ T cells. The transporter associated with antigen processing (TAP) translocates the peptides generated by the proteasome into the ER lumen, where MHC class I molecules fold and assemble. The peptides bind to the MHC class I molecules and the MHC class I-Antigen complex is transported to the plasma membrane [61]. $CD4^+$ cells help to amplify the host immune response by activating effector cells and recruiting additional immune cells to the site of disease [119]. $CD8^+$ cells secrete cytokine and lyse infected cells.

The adaptive immune response is activated after two to three weeks of infection. The interaction between *Mtb* and cells of both the innate and adaptive immune system results in the secretion of chemokines and cytokines, the most important being tumor necrosis factor-α (TNFα), cytokines of the interleukin-1 family (IL-1β, IL-18), IL-12, and IFNγ. The proinflammatory cytokine TNFα combined with IFNγ activates macrophages to produce nitric

oxide synthase (NOS2), allowing the macrophage to kill intracellularly replicating *Mtb*.

However, *Mtb* has evolved various strategies to be able to survive in the hostile environment of the host cells, one of them is phagosome maturation arrest. *Mtb* derived lipid mannose-capped lipoarabinomannan (ManLAM) interferes with Ca fluxes, thereby suppressing hVPS34, necessary for PI3P production [138]. *Mtb* also secretes a lipid phosphatase, SapM, that can dephosphorylate PI3P, therefore blocking phagosome-late endosome fusion [139]. *Mtb* may also inhibit, subvert or modulate antigen presentation by MHC class I, class II and CD1 molecules [10]. For example, *Mtb* disrupts MHC class II presentation in several ways. The prolonged binding of *Mtb* 19 kDa lipoprotein LpqH to TLR2 leads to increased expression of several transcription factors which repress transcription of the MHC class II transcriptional transactivator (CIITA), resulting in decreased MHC II expression on *Mtb* infected cells [144, 10]. Mycobacterial infection also induces the production of interleukin (IL)-10 suppressing cathepsin S, the most important protease capable of mediating the late steps of Ii cleavage needed to generate MHC class II molecules that can be efficiently loaded with peptide antigens [10]. Due to the nature of its cell wall, *Mtb* is also able to resist oxidative radicals, such as nitrogen and oxygen species, intended to kill the pathogen within the macrophages.

1.3.4 *Mtb* and nutrient acquisition

Host defence against pathogens also consists of restricting the availability of nutrients to the pathogen as the pathogen may exploit the host sources for its survival. *Mtb* exploits the host to acquire nutrients. For instance, *Mtb* possesses proteins that are involved in uptake of nutrients such as phosphate, sulfate and some amino acids. Therefore host defence against pathogens consist also of restricting the availability of nutrients to the pathogen as illustrated by iron or tryptophan.

Iron is an indispensable nutrient for almost all organisms [117]. Iron is essential for growth of *Mtb* [110]. In the host, it is required for host metabolism and other important host functions, such as production of reactive oxygen intermediates (ROI) and reactive nitrogen intermediates (RNI) [117]. Iron circulates in the blood stream in the form of a transferrin-iron complex. Macrophages have the transferrin receptor (TfR) which binds the transferrin (Tf)-iron complex. The Tf/TfR complex is then internalised via the classical endocytic pathway into an early endosomal compartment. Once inside the cell, the iron is released from Tf and used by the cell. Unused iron is stored bound to ferritin, the iron storage molecules. Since *Mtb* blocks phagosome maturation at an early endosomal stage, it can access host iron through the Tf/TfR complex [34]. However, IFN-γ-activated macrophages

reduce intracellular iron by downregulating expression of the TfR in order to reduce bacterial growth [117]. To maintain their intracellular iron supply, *Mtb* produce small molecules called siderophores which are secreted into the extracellular space, bind iron and are reinternalised via specific cell surface receptors [34].

Mtb may acquire amino acids from the host and synthesize its own amino acids when they are not available from the host. For instance, *Mtb* is able to import tryptophan from the host, and tryptophan auxotroph are unable to survive single-amino-acid starvation [97]. Upon infection, the host limits tryptophan availability inside the macrophages. This mechanism is activated by IFN-γ, which induces expression of indoleamine 2,3-dioxygenase (IDO) to catabolize transformation of L-tryptophan to N-formylkynurenine, leading to tryptophan depletion inside the cell [117]. Restriction from host-cell sources activates the endogenous tryptophan synthesis of the microbe [117].

The outcome of *Mtb* infection thus depends on the interplay between the host and pathogen. The pathogen proteins interacting with human host might be essential for survival of the pathogen inside the host by acquiring nutrients or interfering with the host defence system. Therefore, knowledge of the set of interacting human and mycobacterial proteins may help us understand the mechanisms of infection and aid in the design of new drugs. Before attempting to find the host-pathogen protein interactions, the protein interaction networks of both the host and the pathogen have to be constructed. The interactions occurring within the same species are known as intra-species interactions, whereas those occurring between different species are called inter-species interactions.

1.4 Protein-protein interactions

Proteins rarely work in isolation, but rather interact with other proteins to carry out their functions. The term interaction does not refer only to physical contact between proteins in a cell but refers to functional interactions. Functional interactions or relationships between proteins involve characteristics which allow particular proteins to carry out their functions together with related proteins without necessarily having physical contact. The set of protein interactions in an organism forms the protein-protein interaction (PPI) network. In this work, we are using all functional interactions for the host and pathogen PPI networks. Therefore, the networks include interactions without direct physical contact. PPIs are detected experimentally or by computational analysis.

1.4.1 Experimental methods for PPIs detection

There are many experimental methods for detecting PPIs. These techniques facilitate a large-scale determination of PPIs. However, these data are prone to high false positive and false negative interaction rates [140]. Co-immunoprecipitations, surface plasmon resonance, nuclear magnetic resonance, X-ray crystallography, label transfer, FRET and far western blot are small-scale experiments for detecting PPIs. Although producing more accurate results compared to high-throughput methods, small-scale experiments are limited in terms of coverage of the interactome. Some of the experimental methods are described below.

- **Yeast two-hybrid (Y2H)** [127]

  The Y2H method is based on transcriptional activation. DNA transcription requires a protein called a transcriptional activator (TA), which contains a DNA-binding domain, for binding to DNA, and an activation domain, for activating transcription of DNA. In order to initiate transcription, the TA binds to the "promoter", a region situated upstream from the gene, via its binding-domain. The activation domain of the TA will then activate transcription. Therefore, transcription of a gene will fail if either of these domains is absent from the TA. In the two-hybrid assay, the protein of interest, the "bait", is fused to a DNA-binding domain by inserting the segment of DNA encoding the bait into a plasmid. Similarly, proteins that bind to the bait, the "fish", are fused to a transcription activation domain. Any protein that binds to the bait will activate the transcription of a reporter gene, which is a gene whose protein product can be easily detected and measured.

- **Mass spectrometry (MS)** [15]

  The MS method is based on the production of ionized peptides which can be detected based on their mass-to-charge ratios, allowing the identification of the original protein or their polypeptide sequences through their masses. MS is often coupled to affinity purification techniques to identify protein-protein interactions. Here we describe the general strategy used to characterize protein complex composition using MS.

  A protein of interest is tagged with a specific tag. Crude sample containing the tagged protein, along with its binding partners, is then incubated with a support where a ligand of the tag in the protein of interest is chemically immobilized. The target protein in the sample will bind to the immobilized ligand. Other sample components are washed away from the support. The target protein and its interactors are then dissociated and recovered from the immobilized ligand. Eluted protein complexes are then separated using techniques such as SDS-page, isoelectric focusing, or various two-

dimensional separation methods. Isolated proteins are proteolytically digested and analyzed by MS. The isolated proteins are converted into ions by using Electrospray Ionization and Matrix Assisted Laser Desorption Ionization (MALDI). These ions are separated according to their mass (m) -to-charge (z) ratios (m/z). The separated ions are detected and this signal is sent to a data system where the m/z ratios are stored together with their relative abundance for presentation in the format of a m/z spectrum.

- **DNA and protein microarrays** [22]

  The technique using DNA microarrays is based on the assumption that interacting proteins are more likely to have their genes co-expressed. Microarrays measure the expression levels of genes by measuring the kinds and amounts of messenger RNA (mRNA) in a cell. DNA microarrays are flat surfaces, typically glass microscope slides, which contain 10,000 to 100,000 spots. Each spot contains multiple single-stranded DNA called the probe and it represents one gene. An mRNA-DNA bond is formed when an mRNA molecule, which is complementary to the probe, hybridizes to that probe. The mRNA molecules, also called targets, are labelled with a fluorescent dye before hybridization. When the targets are exposed to the microarray, the level of fluorescence of the dye will represent the amount of hybridization that has taken place. While DNA microarrays don't measure physical bind of proteins like the methods described above, some protein microarrays, for example antibody arrays, are able to detect physical binding.

Experimental interaction detection is generally expensive, time-consuming and has low accuracy [23]. Computational approaches have therefore been proposed to deal with these problems.

1.4.2 Computational methods for PPIs prediction

Computational methods are divided into two major groups: those that can detect functional links and those that predict physical protein-protein interactions. Methods such as gene neighbour, gene fusion and phylogenetic profile predict functional links while physical interaction is detected by domain-domain interactions or interologs.

- **Gene neighbourhood or co-localization**

  This method is based on the notion that functionally associated genes will stay close to each other on the genome. In bacteria and archae, genes that work together are

generally transcribed as a single unit or operon. In eukaryotes, genes that are observed in close proximity several times across many genomes are predicted to functionally interact [122]. In fact, genes involved in the same biological process or pathway are frequently situated in close genomic proximity.

- **Phylogenetic profile**

  The co-occurence of pairs of genes across multiple genomes suggest that these genes are functionally associated. In this method, a gene is described by its phylogenetic profile which is a record of the absence or presence of the proteins in a given set of genomes. Proteins that have similar phylogenetic profiles are predicted to interact [100].

- **Gene fusion**

  There is a gene fusion event when two separate parent genes are physically fused into a single multifunctional gene. Gene fusion can be viewed as the extreme form of gene co-localization because interacting genes are not only close to each other on the genome, but are physically joined in one single gene. In this approach, one looks for homologous protein sequences to a reference protein (fused protein), but not to each other and which align to different regions of the reference protein. This suggests that the two initial proteins are functionally linked [45].

- **Text mining**

  Text mining methods are used to infer protein-protein interactions from the large amount of biomedical literature. Text mining techniques to derive PPIs are mainly divided into two groups, one using statistical method and the other one using computational linguistic method. Statistical methods infer interactions between proteins by calculating their co-occurrence frequencies. More precisely, the higher the frequencies of two proteins appearing in the same sentences, paragraphs or articles, the more likely these two proteins interact. Computational linguistic methods analyse sentence structures by defining grammars to define sentence structures and using parsers to extract syntactic information and internal dependencies within individual sentences [17].

1.5 Host-pathogen interactions

The experimental methods for detecting intra-species protein interactions can also be applied to prediction of inter-species interactions. In this section, we describe some of the computational methods that have been used for predicting host-pathogen interactions and the databases hosting inter-species interactions.

1.5.1 Computational methods for predicting host-pathogen protein interactions

There are several methods for predicting inter-species protein interactions, most of which have been applied to *Plasmodium falciparum* and HIV protein interaction with human. We briefly introduce some of these methods here.

Dyer *et al.* [43] predicted human-*Plasmodium falciparum* protein interactions by integrating known intra-species PPIs with protein-domain profiles. They first identified the functional domains in each interacting protein of a set of intra-species PPIs. Then, using Bayesian statistics, they assess the probability that two proteins with a pair of domains will interact, for every pair of functional domains. They predicted 516 PPIs between the two organisms using this method.

Lee *et al.* [77] used interologs to predict human-*Plasmodium falciparum* protein interactions. Interologs are conserved protein interactions between organisms. In this method, two host-pathogen proteins are predicted to interact if they have interacting orthologs in another organism. They inferred more than 3,000 human-*Plasmodium falciparum* PPIs. Since not all the interactions are likely to take place, they used gene ontology annotations to filter the interactions, leading to 918 host-pathogen interactions.

Interaction prediction between HIV and human proteins has also been performed. Tastan *et al.* [131] used a Random Forest classifier (RF) to classify protein pairs as either "interacting" or "not-interacting" based on biological information sources. Evans *et al.* [46] predicted HIV-human protein interactions using virus and host sequence motifs. Doolittle and Gomez [39] predicted interactions between HIV and human proteins based on structural similarity of HIV proteins to human proteins having known interactions. In the context of *Mtb*-human protein interactions, Krishnadev *et al.* [72] applied a homology-based method to predict protein interactions. Another study by Davis *et al.* [37] used comparative modelling to predict host-pathogen PPIs with ten pathogens, including *Mycobacterium tuberculosis*. From host and pathogen protein pairs sharing similarity to protein complexes with known structures, 3D structural models of putative complexes were built.

1.5.2 Databases of host-pathogen interactions

Although many databases store intra-species interactions, there is a limited number of databases designed specifically for inter-species interactions. Moreover, they are limited in terms of species coverage. Some of these databases are described below.

- **Host-Pathogen Interaction database** (http://agbase.msstate.edu/hpi/main.html) [75]

  Host-Pathogen Interaction database (HPIDB) integrates experimental PPIs from various public databases and also allows the retrieval of homologous host/pathogen sequences. HPIDB contains interactions between 49 hosts and 319 pathogens. Host species include human, mouse, mouse-ear cress, rat, bovine and chicken. Pathogenic species include bacteria, virus, protist and fungi.

- **Pathosystems Resource Integration Center** (http://www.patricbrc.org/)

  Pathosystems Resource Integration Center or PATRIC is an information system integrating a comprehensive bacterial genomics database with computational tools for bioinformatics analysis. PATRIC contains experimentally confirmed protein interactions between human and bacterial proteins.

- **Pathogen-Host Interaction Data Integration and Analysis System** (http://www.phidias.us)

  Pathogen-Host Interaction Data Integration and Analysis System (PHIDIAS) allows the search, comparison, and analysis of integrated genome sequences, conserved domains, and gene expression data related to pathogen-host interactions [145]. It includes 42 pathogens, most of them in the category A, B and C priority pathogens identified by the National Institute of Allergy and Infectious Diseases (NIAID) and the Centers for Disease Control and Prevention (CDC) in the USA. The pathogen species also include the human immunodeficiency virus (HIV) and *Plasmodium falciparum*, both of them considered as high priority with regard to public health.

- **Virus–Host Network** (http://pbildb1.univ-lyon1.fr/virhostnet/)

  Virus–Host Network (VirHostNet) database is specialized in virus–virus, virus–host and host–host interaction networks [93].

To our knowledge, there are no databases that provide a comprehensive set of human-*Mycobacterium tuberculosis* protein interactions.

1.6 Network measures and network topology

Biological networks are modelled using graphs, which are a collection of nodes and edges joining two nodes. Graphs are mathematical concepts that give insight into the organisation and structure of complex networks such as protein-protein interaction networks. In this

section, we first present some network measures that are used to analyse networks and assess their characteristics, thereafter we present the network topologies.

1.6.1 Network measures

There are several network measures used to compare and analyse complex networks.

Degree (or connectivity) of a node
The degree k of a node is the simplest measure characterizing a node and is just the number of links a node has.

Mean path length
The shortest path length between two nodes is the minimum number of edges needed to connect one node to the other. The mean path length is the average path length of the shortest paths between all pairs of nodes in the network. This measure reflects the ability of two nodes to communicate with each other.

Degree distribution
The degree distribution $P(k)$ is the probability that a given node has degree k. $P(k)$ is given by

$$P(k) = \frac{N(k)}{N}$$

where $N(k)$ is the number of nodes with degree k, $k = 1, 2, \ldots$ and N is the total number of nodes in the network. Networks have different topologies according to their degree distribution.

1.6.2 Network topologies

The form of the degree distribution $P(k)$ provides an insight into the network topology, which enables us to have a broad view of the network's characteristics. Here we give two network topologies that are relevant to the understanding of biological networks.

Random network
Random networks have a degree distribution $P(k)$ essentially following a Poisson distribution. This type of network is homogeneous, meaning that most nodes have roughly the same number of links.

Scale-free network
Scale-free networks have a degree distribution following a power law. Scale-free networks are heterogeneous, meaning that they have few nodes with many interactions and

many nodes with few interactions. Several studies have shown that biological networks are scale-free. For instance, investigation of the protein networks of *S. cerevisae* [147], *E. coli*, *D. melanogaster*, *C. elegans* and *H. pylori* showed that they all are scale-free [55].

Analysis of the topological properties of networks is important for drug discovery. In fact, biological networks, which are scale-free, are very resistant to random removal of nodes but rapidly disintegrate when facing targeted removal of the highly connected nodes or hubs, therefore diminishing the ability of the nodes to communicate with other nodes. Jeong *et al.* [3] showed that hubs tend to be essential proteins (proteins that, when knocked out, render the cell unviable).

1.7 Problem statement and overview of thesis

Protein interactions between the host and *Mtb* are essential for *Mtb* survival in the host by modulating the host response to bacterial infection or by acquiring nutrients it requires for its growth. Therefore, the identification of the protein interactions that *Mtb* uses to invade the host can contribute to the process of identifying potential targets for designing new drugs. Unfortunately, experimental studies of host pathogen interactions are very scarce, hence we have to rely on computational methods to predict host-pathogen interactions. Host-pathogen interactions have been predicted using protein-domain profiles, sequence motifs, Random Forest classifier, comparative modelling and interologs. The aim of this project is to predict human-*Mtb* protein-protein interactions using the interologs method with intra-species and inter-species interactions as data for interaction prediction and overlay them onto a human and *Mtb* protein interaction network.

Chapter 2 describes the construction of both the human and *Mtb* PPIs and the prediction of host-pathogen interactions. In Chapter 3, we analyse features of the predicted interactions such as their subcellular locations, pathways and GO biological processes. Figure 1.1 provides a flowchart of the procedure followed in Chapters 2 and 3 to achieve the aims of the project. Finally, Chapter 4 concludes the thesis.

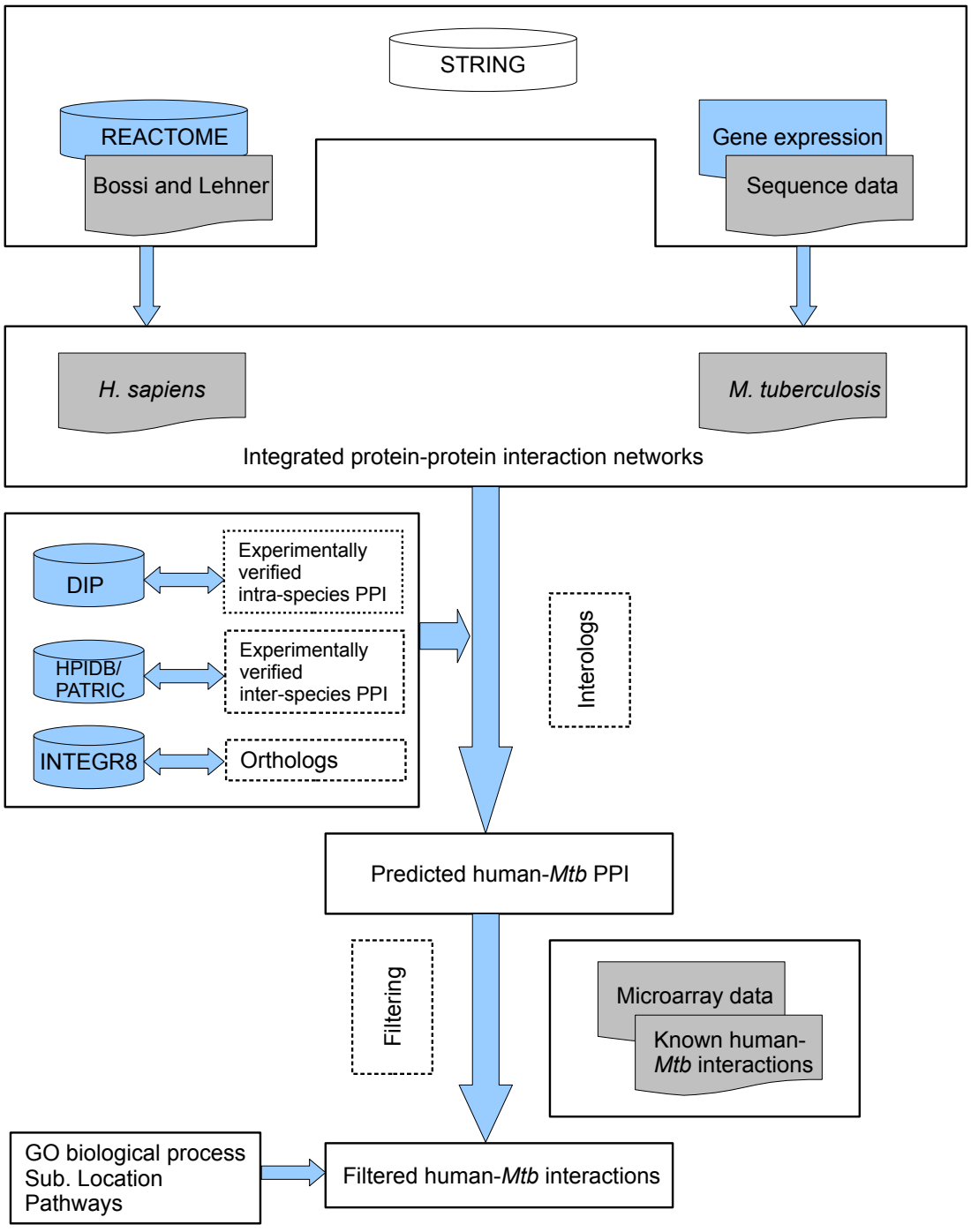


Figure 1.1: Flowchart depicting the prediction of host-pathogen interactions and their analysis

Chapter 2

Network generation and host-pathogen protein-protein interaction prediction

2.1 Human protein-protein interaction network

In 2000, Uetz *et al.* undertook the first large-scale protein interaction study, which was done in yeast [133], and studies have also been done in the fly [53] and the worm [79]. For the human interactome, attempts to map the protein interaction network have also been made by different groups: Bader *et al.* [9], Peri *et al.* [101], and Ramani *et al.* [109] built maps from literature searches, Lehner and Fraser [78], Brown and Jurisica [21], and Persico *et al.* [102] used orthologs, and Rual *et al.* [114] and Stelzl *et al.* (2005) [125] generated their map based on large-scale yeast two-hybrid data.

The human interactome is estimated to contain approximately 20,000-25,000 proteins and roughly 154,000-369,000 interactions. The large variation in the number of interactions is largely due to the methods for predicting these interactions.

Major databases that store human protein-protein interactions are described below.

- **BIND** or Biomolecular Interaction Network Database (http://www.bind.ca)

  BIND consists of biomolecular interaction, reaction, complex and pathway information. Interactions are obtained from curation of published experimental research. BIND contains over 200,000 binary interactions and over 3,700 biological complexes, from more than 1,500 species.

- **BioGRID** or Biological General Repository for Interaction Datasets (http://thebiogrid.org)

  BioGRID is a database of protein and genetic interactions from major model organism species. Data in BioGRID come from curation of large- and small-scale experimental data found in primary literature. The current release of BioGRID (3.1.93) contains 375,704 unique interactions and 44,908 proteins. 40 species are represented in BioGRID, including *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, *Hepatitus C Virus*, *Human Immunodeficiency Virus 1* and 2, *Mus musculus*, *Oryza sativa*, *Plasmodium falciparum 3D7* and *Saccharomyces cerevisiae*. There are 83,592 unique interactions and 15,549 proteins for human.

- **DIP** or Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu)

  DIP collect experimentally determined protein-protein interactions that are curated, both manually by expert curators and also automatically using computational approaches. DIP currently contains 25,023 proteins, 74,355 interactions and 517 organisms. Some selected organisms are *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, *Mus musculus*, *Bos taurus*, *Rattus norvegicus*, *Helicobacter pylori* and *Saccharomyces cerevisiae*. DIP now contains 3,386 human proteins and 1,425 interactions.

- **HPRD** or Human Protein Reference Database (http://www.hprd.org)

  HPRD is dedicated to human proteins. Apart from PPIs, it also contains information about post-translational modifications, subcellular localization, protein domain architecture, tissue expression and association with human diseases. HPRD currently has 30,047 protein entries and 41,327 protein interactions.

- **IntAct** (http://www.ebi.ac.uk/intact)

  Data in IntAct are from the literature or from direct data depositions by expert curators.IntAct contains 62,450 proteins and 204,629 interactions and complexes from more than 275 organisms, though they are focusing on model organisms such as *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, *A. thaliana* and *E. coli*. 89,519 human interactions and more than 17,500 human proteins are now included in IntAct.

- **MINT** or Molecular INteraction Database (http://mint.bio.uniroma2.it/mint/Welcome.do)

  MINT stores interactions between biological molecules and focuses on experimentally verified protein-protein interactions. MINT contains a total of 35,430 proteins and

241,373 interactions from more than 300 organisms. Among them, 8,639 proteins and 26,761 interactions are for human.

2.1.1 Human network generation

We used three datasets to construct our human interaction network. The first one is the human protein interaction network used by Bossi and Lehner [18], the second and the third were downloaded from REACTOME and STRING, respectively.

Bossi and Lehner integrated data from 21 sources to build their human protein network. They only included interactions supported by at least one piece of direct experimental evidence demonstrating physical association between two human proteins. In addition to the direct evidence, some interactions also have supporting evidence. Bossi and Lehner used 21 sources for the data divided into 26 categories. In order to quantify the interactions, each interaction has been assigned a score which depends on the reliability of the source, experimentally verified interactions are given a higher score. The interaction categories and their corresponding scores are shown in Table 2.1.

REACTOME is an expert-curated and highly reliable database of human biological pathways and reactions. The interactions from REACTOME were given a score of 0.95 like the REACTOME source in the Bossi and Lehner data. The REACTOME data used here is more recent data than that used by Bossi and Lehner.

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a database for the exploration and analysis of protein-protein interactions. Data contained in the STRING database are derived from experimental data, as well as from public literature collections and computational prediction methods based on domain fusion, phylogenetic profiling, homology and gene neighbourhood concepts. In STRING, a confidence score is assigned to each identified protein-protein association, derived by benchmarking the performance of the predictions against a common reference set. To provide a unified view of the data, STRING uses a scoring scheme that integrates all the different methods to produce a final "combined score" between any pairwise predicted protein. Assuming that all the methods are independent, the final score S_{STRING} is given by:

$$S_{STRING} = 1 - \prod_{i=1}^{n}(1 - s_i) \tag{2.1}$$

where s_i is the score from the method i and n is the number of methods. An interaction between two proteins has a higher score when it is supported by several types of evidence, expressing an increased confidence.

Source	Description	Score
ATAXIA	Literature, stringent two-hybrid	0.85
CORUM	Experimentally verified mammalian protein complexes	0.9
HPRD	Manual curation of PPIs from literature, experiments in vivo, in vitro and Y2H	0.85
INTACT	Literature curation or direct submissions	0.85
MDC	Y2H plus verification	0.85
OPHID	Data integration, known and predicted PPIs, orthology	0.75
UNILEVER	Y2H, mass spectrometry, literature	0.85
INTNET	PPI predictions from data integration	0.75
BIOVERSE	Data integration	0.75
BIOGRID	Literature curated interactions	0.85
REACTOME	Curated interactions	0.95
BIND	Literature curated interactions	0.85
CCSB-curated	Stringent Y2H	0.78
DIP	Experimentally determined PPIs	0.85
MIPS	Literature curated experimental interactions	0.85
EWING	Affinity purification, mass spectrometry + verification	0.85
JERONIMO	Experimental protein complexes	0.9
HOMOMINT	Human orthologs of experimentally verified PPIs in model organisms	0.7
COCIT	Co-citation	0.65
Phenotype similarity	Share phenotype	0.65
Genetic	Genetic interaction	0.6
Co-expression	Co-expression	0.65
Domain-domain interaction	Domain-domain interaction	0.65
Shared GO annotation	Share GO process	0.5
Interolog	Inferred from orthologs	0.7
Share upstream motif	Shared upstream region	0.5

Table 2.1: Sources and scores for the Bossi and Lehner data.

All the three networks above have different IDs for the genes, so we converted all the IDs to UniProt IDs in order to unify the data. We used the database identifier mapping (ID Mapping) tool in UniProt for this purpose. This tool maps a lists of identifiers from an external database to UniProtKB, taking as input database identifiers. In our case, we uploaded the files containing the nodes of our three networks. As the source database, we selected Ensembl and Ensembl protein for the Bossi and Lehner and STRING networks, respectively. The REACTOME network already had UniProt IDs. The target database was set to UniProtKB AC. We then downloaded tab-separated tables mapping the source database identifiers to UniProt identifier. After converting the IDs of the two networks to UniProt IDs, we combined the three individual networks to form a single network, where self-interacting proteins were removed. Each interaction of this network was represented by the interacting proteins, along with the various scores from the individual data, including the 26 scores from the Bossi and Lehner network, the REACTOME score and the final STRING score. The same formula as in Equation (2.1) was used to compute the final score S for each interaction, but now s_i is the score from each different source and n is the number of sources. Therefore, an interaction predicted by more than one database has a higher score, expressing an increased reliability. We obtained a human network consisting of 18,343 proteins and 1,339,048 interactions. To increase the reliability of the network, only interactions that have a final score $S \geq 0.65$ were used to build the network. A cut-off score of 0.65 was chosen because the interactions in the Bossi and Lehner network which have a score greater or equal to 0.65 are considered to be reliable. 1,004,978 interactions and 1,795 proteins were filtered out during this process. Ultimately, we obtained a human network of 16,548 proteins and 334,070 interactions.

2.1.2 Topological properties of the human PPI network

In order to analyse and manipulate the netwoks, we used `NetworkX` which is a library for the Python scripting language. `NetworkX` allows the creation, manipulation, and study of the structure and dynamics of complex networks (`http://networkx.lanl.gov/`). To study the topological properties of our network, we created a tab-delimited file of the graph, where a line is an interaction between two proteins represented by their UniProt accession. We then computed the characteristics of the network and results are shown in Table 2.2.

Proteins in the human network possess an average of 40 neighbours. As seen in Figure 2.1, the node degree distribution of the human network shows a scale-free network topological property, where few proteins are highly connected and most proteins have a low number of interactions. The average shortest path length of 3.507 is small compared to the network's

Number of proteins	16,548
Number of interactions	334,070
Average degree	40.3759
Number of connected components	52
Percentage of nodes in the largest connected component	99.35%
Average shortest path length	3.507646

Table 2.2: Characteristics of the human network

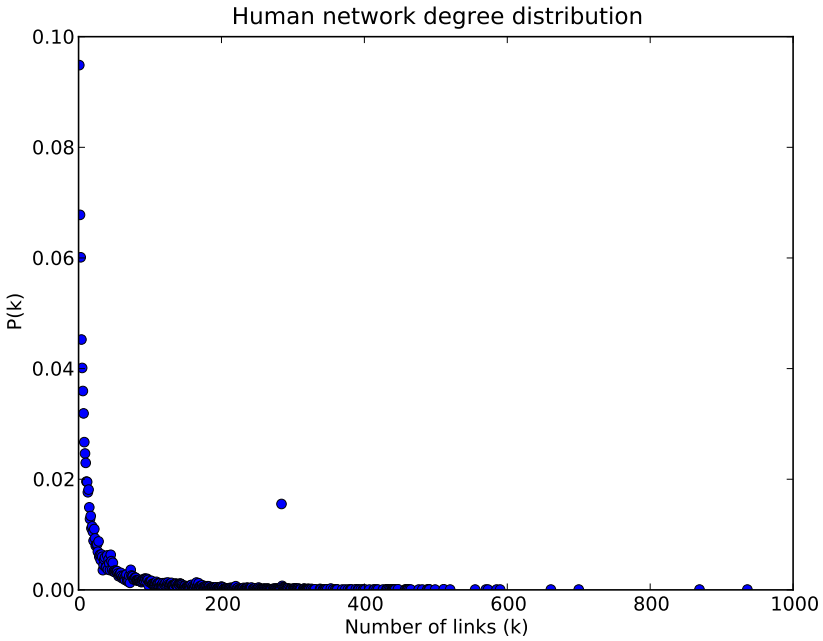


Figure 2.1: Human network degree distribution

size, showing that our human network has a "small world property" – a path of a few links can connect any two nodes in the network.

The node degree distribution of the human network shows an outlier. These are 257 proteins of degree 284. These proteins are olfactory receptors except Q92620, which is a pre-mRNA-splicing factor ATP-dependent RNA helicase PRP16. The olfactory receptor proteins form a clique (every two nodes connected by an edge) and are connected to the same 29 proteins. All the interactions involving the 257 proteins come from STRING.

2.2 Mtb protein-protein interaction network and properties

Yeast-two hybrid (Y2H) systems have been used to investigate protein-protein interaction networks in a number of models including human, *C. elegans* and *D. melanogaster*. However, very few studies have been done on bacterial species. For *Mtb* in particular, Wang *et al.* [142] constructed a global *Mtb* network using yeast-two hybrid. Their network contains 8,042 interactions between 2,907 proteins, representing about 74.1 % of functional proteins encoded by the entire genome. The STRING database (described in section 2.1.1) contains PPI networks for various *Mtb* strains namely *Mycobacterium tuberculosis* CDC1551, *Mycobacterium tuberculosis* H37Rv, *Mycobacterium tuberculosis* F11, *Mycobacterium tuberculosis* H37Ra, and *Mycobacterium tuberculosis* KZN 1435.

2.2.1 *Mtb* network generation

The *Mycobacterium tuberculosis* genome contains roughly 4,000 protein coding genes. The protein interaction network of *Mtb* strain CDC1551 was previously generated through a combination of data extracted from the STRING database and from gene expression and sequence data [84]. From STRING, the following prediction methods were used: neighbourhood, gene fusion, co-occurence, experiments, databases and text mining. Interactions from gene expression data were derived from multiple microarray datasets and use a random partial least squares approach for scoring interactions [87]. Interactions from sequence data comprise sequence similarity and domain data. The similarity score used here comes from [85] and is defined by:

$$R(s_1, s_2) = \frac{S(s_1, s_2) + S(s_2, s_1)}{2\max\{S(s_1, s_1), S(s_2, s_2)\}}$$

where $S(s_i, s_j)$ is the bit score alignment of homologous sequences s_i and s_j.

The relationship score between pairwise proteins i and j sharing common InterPro hits is given by:

$$r_{ij}(\delta) = (1 - H_2(\delta))/bit$$

where $H_2(\delta)$ is the binary entropy function measuring the uncertainty related to the number n of common InterPro hits in the set of InterPro hits. This entropy is given by

$$H_2(\delta) = -\delta \log_2(\delta) - (1 - \delta) \log_2(1 - \delta)$$

with $\delta \equiv \delta(n, \sigma, \alpha) = \phi\left(\frac{n^\alpha}{\sigma}\right)$. The function ϕ is the cumulative probability of the standard

Gaussian distribution defined by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} exp\left(-\frac{x^2}{2}\right) dx$$

where σ denotes the standard deviation and $\alpha \geq 0.05$, the calibration control parameter, strengthening the impact of the confidence-level for the data under consideration [85].

The same scoring method used for the human network was used to unify the *Mtb* data. The interactions in STRING are divided into a confidence range: low confidence if the $score < 0.3$, medium confidence if $0.3 \leq score \leq 0.7$ and high confidence if the $score > 0.7$. A cut-off of 0.5 was set to select the interactions that are in the final network used for the analysis so that we would not pick interactions with a score too low but still have medium confidence interactions. In the end, we obtained 4,070 proteins joined by 38,049 edges.

2.2.2 Topological properties of the *Mtb* network

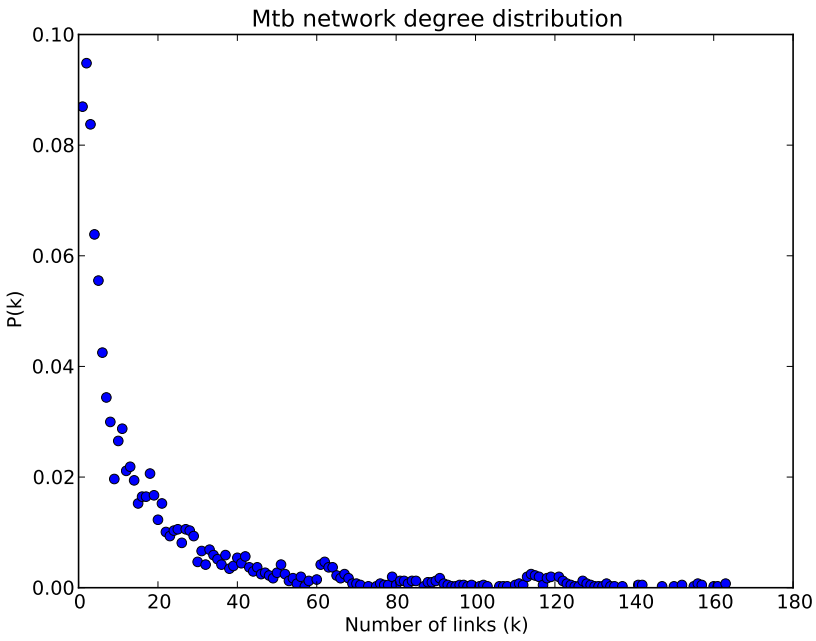


Figure 2.2: Degree distribution of the *Mtb* network

The characteristics of the final *Mtb* network are shown in Table 2.3. Our *Mtb* network has 68 connected components, where 95.97% of the nodes belong to the largest connected component. The average degree of a node in the *Mtb* network is 18.69. Like the human network the *Mtb* network has a "small-world property" indicated by a very low average

Number of proteins	4,070
Number of interactions	38,049
Average degree	18.693
Number of connected components	68
Percentage of nodes in the largest connected component	95.97%
Average shortest path length	4.241

Table 2.3: Characteristics of the *Mtb* network

shortest path length of 4.24 and the degree distribution also shows a scale-free network (Figure 2.2).

2.3 Inter-species host-pathogen interaction prediction

2.3.1 Known host-pathogen interactions

Known human-*Mtb* protein interactions are not stored in databases as is the case for many human-virus protein interactions. Therefore, manual curation of existing literature had to be done in order to gather as many host-pathogen protein interactions as possible. Through literature mining, we were able to retrieve 47 inter-species interactions between *Mtb* and human where both interacting proteins were present in the networks we built. A list of these known interactions is provided in Table 2.4.

The known human-*Mtb* protein interactions are mostly those involved in the recognition of the pathogen by the host. Specific host receptors recognize specific *Mtb* components, for example, the toll-like receptors (TLRs) recognize various bacterial lipoproteins (lprA [99], lpqH [24], lprG [52]) whereas the human pulmonary surfactant protein A (PSP-A) binds to *Mtb* surface glycoproteins [107]. The human plasminogen can bind to 15 different bacterial plasminogen receptors [146].

Mtb protein UniProt Acc	*Mtb* protein name	Human protein UniProt Acc	Human protein name	Reference
P0A5Q4	Immunogenic protein MPT64	P25685	DnaJ homolog subfamily B member 1	[32]
P0A5J0	Lipoprotein lpqH	O60603	Toll-like receptor 2	[24]
P0A520	60 kDa chaperonin 2	O00206	Toll-like receptor 4	[24]

P0A5B9	Chaperone protein DnaK	O00206	Toll-like receptor 4	[24]
P0A5B9	Chaperone protein DnaK	O60603	Toll-like receptor 2	[24]
P0A4V2	Antigen 85-A	P02751	Fibronectin	[98]
P0C5B9	Antigen 85-B	P02751	Fibronectin	[98]
P0A4V4	Antigen 85-C	P02751	Fibronectin	[98]
Q7D8M9	PPE family protein, PPE18	O60603	Toll-like receptor 2	[92]
Q50615	Uncharacterized PE-PGRS family protein PE_PGRS33	O60603	Toll-like receptor 2	[14]
Q11049	Putative lipoprotein lprA	O60603	Toll-like receptor 2	[99]
P0A5I8	Lipoprotein lprG	O60603	Toll-like receptor 2	[52]
P15712	Phosphate-binding protein pstS 1	O60603	Toll-like receptor 2	[67]
P15712	Phosphate-binding protein pstS 1	O00206	Toll-like receptor 4	[67]
Q50906	Alanine and proline-rich secreted protein apa	Q9NNX6	CD209 antigen	[104]
P0A5J0	Lipoprotein lpqH	Q9NNX6	CD209 antigen	[104]
P0A5B9	Chaperone protein DnaK	P00747	Plasminogen	[146]
P0A590	Glutamine synthetase 1	P00747	Plasminogen	[146]
P60176	Adenosylhomocysteinase	P00747	Plasminogen	[146]
P66004	Dihydrolipoyl dehydrogenase	P00747	Plasminogen	[146]
P0A5H3	Isocitrate lyase	P00747	Plasminogen	[146]
P0A558	Elongation factor Tu	P00747	Plasminogen	[146]
P77899	S-adenosylmethionine synthase	P00747	Plasminogen	[146]
P67475	Fructose-bisphosphate aldolase	P00747	Plasminogen	[146]
P0A4V2	Antigen 85-A	P00747	Plasminogen	[146]
P0C5B9	Antigen 85-B	P00747	Plasminogen	[146]
P0A4V4	Antigen 85-C	P00747	Plasminogen	[146]
P0A4V6	MPT51/MPB51 antigen	P00747	Plasminogen	[146]
O33245	Proteasome subunit beta	P00747	Plasminogen	[146]
P0A5Q4	Immunogenic protein MPT64	P00747	Plasminogen	[146]
P09621	10 kDa chaperonin	P00747	Plasminogen	[146]
P0A5B9	Chaperone protein DnaK	Q9NNX6	CD209 antigen	[30]
P0A518	60 kDa chaperonin 1	Q9NNX6	CD209 antigen	[30]
P0A5I8	Lipoprotein lprG	Q9NNX6	CD209 antigen	[30]
P0A5I8	Lipoprotein lprG	Q9H2X3	C-type lectin domain family 4 member M	[30]
P0A4V4	Antigen 85-C	P05107	Integrin beta-2	[60]
P0A5J0	Lipoprotein lpqH	Q15399	Toll-like receptor 1	[40]

P0A5J0	Lipoprotein lpqH	P08571	Monocyte differentiation antigen CD14	[40]
Q11049	Putative lipoprotein lprA	Q15399	Toll-like receptor 1	[40]
Q11049	Putative lipoprotein lprA	Q9Y2C9	Toll-like receptor 6	[40]
Q11049	Putative lipoprotein lprA	P08571	Monocyte differentiation antigen CD14	[40]
P0A5I8	Lipoprotein lprG	Q15399	Toll-like receptor 1	[40]
P0A5I8	Lipoprotein lprG	P08571	Monocyte differentiation antigen CD14	[40]
P15712	Phosphate-binding protein pstS 1	Q15399	Toll-like receptor 1	[40]
Q50906	Alanine and proline-rich secreted protein apa	Q8IWL1	Pulmonary surfactant-associated protein A2	[107]
Q50906	Alanine and proline-rich secreted protein apa	Q8IWL2	Pulmonary surfactant-associated protein A1	[107]
P0A564	6 kDa early secretory antigenic target	O00560	Syntenin-1	[120]

Table 2.4: Known host-pathogen interactions from the literature

2.3.2 Host-pathogen interaction prediction using interologs

The term "interolog" was first introduced by Walhout *et al.* [141]. Interologs are conserved interactions between a pair of proteins which have interacting orthologs in another organism. More precisely, the interaction X/Y in one species is referred to as interologs of X'/Y' in another species if X' and Y' are orthologs of X and Y, respectively.

The interolog method has mostly been used to infer intra-species interactions in higher order organisms such as *Homo sapiens* and has also been applied to inter-species interaction prediction. For example, Lee *et al.* used interologs to predict host-pathogen interactions between *Homo sapiens* and *Plasmodium falciparum* [77]. An interaction between a *P. falciparum* protein P_a and a human protein H_b is inferred if the ortholog S_a of P_a and the ortholog S_b of H_b in species S interact. We used the same principle to infer interactions between *Homo sapiens* and *Mtb*. The human and *Mtb* orthologs were downloaded from INTEGR8 (`http://www.ebi.ac.uk/integr8`), and the known interactions were extracted from the DIP database. To infer an interaction between a human protein and a *Mtb* protein, we first took a human protein H_a and determined whether it had an ortholog S_a. If this was the case, then we looked for the interacting partner S_b of S_a in the DIP database. If an interacting partner S_b was found, we checked whether it had an ortholog M_b in *Mtb*. When all the requirements were fulfilled, we inferred that H_a and M_b are interologs. This process

is depicted in Figure 2.3. We applied the same process again, but starting with the *Mtb* proteins.

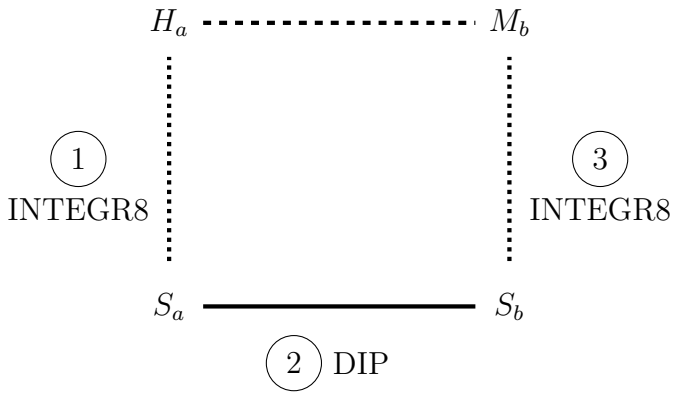


Figure 2.3: Steps for inferring inter-species interactions using the interologs method. Dotted lines represent orthologs, the solid line represents intra-species interaction and the dashed line represents inferred inter-species interaction.

A total of 483 host-pathogen interactions between 175 human and 192 *Mtb* proteins were found using the interolog method described above. 10 reference species were used to infer the interologs. Table 2.5 lists the reference species along with the resulting number of interologs. The original species are phylogenetically close enough to either human or *Mtb* to assume no false positives and we assumed that orthologs should have conserved functions.

Original species	Number of interactions
M. smegmatis	1
M. musculus	10
M. pneumoniae M129	65
B. subtilis	2
H. pylori	11
D. melanogaster	28
C. elegans	46
E. coli ATCC 27325	54
E. coli MG1655	260
R. norvegicus	6

Table 2.5: Original species and number of interactions for the interologs

The interactions were further investigated by filtering based on two criteria: connections to known interactions and whether they are expressed during infection.

2.3.3 Filtering for "interolog-known" interactions

In order to analyse the interologs predicted, we first looked at the cases where both proteins were neighbours of known interactions, referred to as "interolog-known" interactions. The network definition of neighbourhood is used here, which is two proteins are neighbours if there exists an edge connecting them in the network. Practically, to predict the "interolog-known" interactions, we took all possible `(interolog,known)` pairs where `interolog` is an inter-species interaction predicted by interologs and `known` is a known inter-species interaction. If the human interolog protein is a neighbour of the human known protein and the *Mtb* interolog protein is a neighbour of the *Mtb* known protein, then `interolog` is an "interolog-known" interaction. Among the 483 host-pathogen interactions, 3 interactions fall into the "interolog-known" category. We predicted that the human P07814 protein, which is a bifunctional aminoacyl-tRNA synthetase, interacts with P0A5U4 (RecA, recombinase A) from *Mtb*. The RecA protein is a recombinase functioning in recombinational DNA repair in bacteria. These two proteins are the neighbours of the human protein P00747 (plasminogen) and the mycobacterial protein P77899 (S-adenosylmethionine synthase), respectively. An interaction between the human protein ATP synthase subunit beta, mitochondrial, P06576 and the *Mtb* protein P0A548, a chaperone protein DnaJ 1, was also predicted. This second predicted interaction is the neighbour of three known interactions between one human protein (P00747) and three *Mtb* proteins: P0A5B9 (chaperone protein DnaK), P0A558 (elongation factor Tu) and P77899. The third interaction involves the human protein P27361, which is a mitogen-activated protein kinase 3 (MAPK3), and the probable acetyl-CoA acyltransferase fadA2 from *Mtb*. MAPK3 is an enzyme which is a member of a MAPK family. Induction of the MAPK pathway is required for the expression of TNF-alpha, IL-10, and MCP-1 by human monocytes during *M. tuberculosis* H37Rv infection [124]. The third interaction is the neighbour of the interaction between the human protein O00206 (toll-like receptor 4) and the *Mtb* protein P0A520 (60 kDa chaperonin 2). The "interolog-known" interactions are listed in Table 2.6 and are visualised in Figure 2.4 together with the known interactions of which they are the neighbours. Network visualisation was done using Cytoscape (`http://www.cytoscape.org/`), which is a popular bioinformatics package for biological network visualization and data integration. To create a network in Cytoscape, we imported tab-delimited text file of the network. One row in the file represented an edge and its edge attribute. In our case, the edge attributes are the type of the interactions, such as inter-species or intra-species interactions.

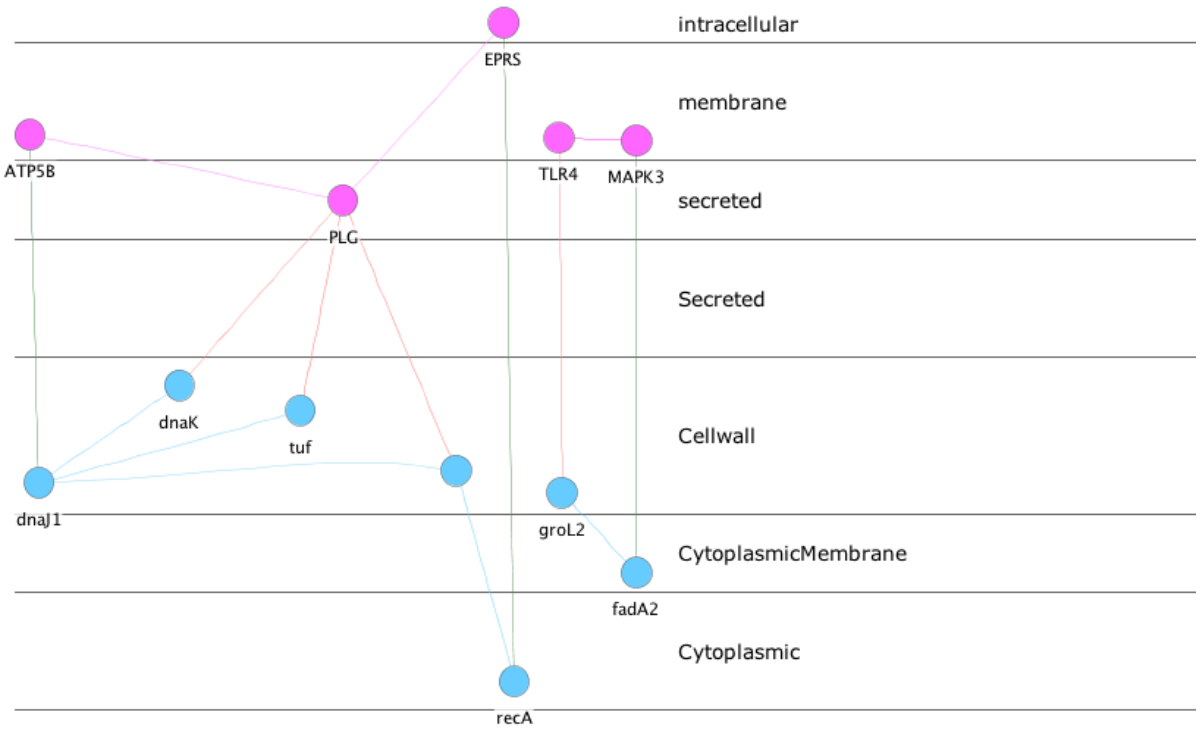


Figure 2.4: Interactions predicted by interologs which are neighbour of known interactions. Pink nodes are human proteins and pink edges are human PPIs. Blue nodes are *Mtb* proteins are blue edges are *Mtb* PPIs. Red edges are known human-*Mtb* interactions and green edges are inter-species interactions predicted by interologs.

Human UniProt ID	Human protein name	*Mtb* UniProt ID	*Mtb* protein name
P07814	EPRS, bifunctional aminoacyl-tRNA synthetase	P0A5U4	RecA, recombinase A
P06576	ATP5B, ATP synthase subunit beta, mitochondrial	P0A548	DnaJ1, chaperone protein DnaJ 1
P27361	MAPK3, mitogen-activated protein kinase 3	O86361	fadA2, probable acetyl-CoA acyltransferase FADA2

Table 2.6: The "interolog-known" interactions

2.3.4 Filtering for "interolog-array" interactions

We also looked at the interologs where both of the proteins are differentially expressed in microarray data, which we shall refer to as "interolog-array" interactions.

The outcome of an infection depends on how the host responds to the pathogen and how the pathogen evades the immune system. Investigation of the entire transcriptome of a cell during infection may provide a hint about the host-pathogen cross-talk. This is easily feasible using microarray technology which allows the simultaneous analysis of expression of thousand of genes. Therefore, this method has been widely used to study the interplay between the host and the pathogen. Unfortunately, the microarray data gives us only set of genes that might interact as they are expressed under the appropriate conditions but do not tell us explicitly which gene interacts with which other gene. That is why we combined the interactions predicted by interologs with the microarray data: the interologs give the interactions and the microarray data ensure that the interactions could occur.

For human-*Mtb* interactions, a number of groups have studied the global gene expression profile of *Mtb*-infected human macrophages (Mφs) and dendritic cells (DCs) [31, 108], and some analysed the mycobacterial transcriptome in human Mφs [28] or in human lung tissue samples [106]. These studies have been conducted in either the host or the pathogen. We used microarray data from Tailleux *et al.* for our analysis [128]. They used microarray technology to decipher transcriptional changes upon infection both in *Mtb* and in Mφs and DCs derived from the same donors [128] over a time-course experiment. Here, we describe briefly the procedures carried out and data analysis of the experiment, a detailed description can be found in [128].

Cellular RNA was extracted from 9 individual donors at 4, 18 and 48 hours of infection and at the time of infection (reference), producing a total of 72 samples. The samples were labelled and hybridised to Human U133A oligonucleotide microarray chips with 22,283 probe sets according to the manufacturer's protocols. The arrays were then scanned and analysed.

They applied a filter based on Detection calls to filter out noisy data before selecting differentially expressed genes: they first removed the probe sets called "Absent" over all conditions and replicates; then they determined the 95^{th} percentile of all the signals of the entire dataset flagged with an absent call and used it as a threshold to remove all the remaining probe sets whose expression values were always below the threshold in each sample. The remaining probe sets were used for the analysis.

Differentially expressed genes were detected using the Limma Bioconductor library, which is based on the fitting of a linear model to estimate the variability in the data. A threshold p-value of 10^{-4} was used to select differentially expressed genes.

Mycobacterial RNA extracted from infected Mφs and DCs from 3 healthy donors at 1, 4 and 18 hours after infection, and from two biological replicates of log phase *in vitro* growth were hybridised in duplicate to a *M. tuberculosis* whole genome microarray. The hybridised slides were then scanned sequentially and the data was analysed using the Limma software package.

Raw human microarray data used in our analysis was provided by the authors of the article. We performed the analysis described above on the data to obtain the human genes differentially expressed during *Mtb* infection. On the other hand, the list of *Mtb* genes differentially expressed were downloaded from BμG@Sbase under the accession E-BUGS-58 (http://bugs.sgul.ac.uk/bugsbase).

To find the "interolog-array" interactions, we took the interologs and found those where both human and *Mtb* proteins were differentially expressed in the above experiments. This led to 78 interactions between 35 human proteins and 47 *Mtb* proteins shown in Table 2.7.

Human Acc	Uniprot	Human protein name	*Mtb* Acc	Uniprot	*Mtb* protein name
Q12965		MYO1E, myosin-Ie	P67510		LeuS, leucyl-tRNA synthetase
Q9P2R7		SUCLA2, succinyl-CoA ligase [ADP-forming] subunit beta, mitochondrial	P96377		Eno, enolase
			P71558		SucD, succinyl-CoA ligase [ADP-forming] subunit alpha
			O06249		GadB, glutamate decarboxylase
			P63288		ClpB, chaperone protein ClpB
			O06147		RpsA, 30S ribosomal protein S1
			P0A658		Upp, uracil phosphoribosyltransferase
			P0A5U4		RecA, recombinase A

P60891	PRPS1, ribose-phosphate pyrophosphokinase 1	P0A5X6	RpsC, 30S ribosomal protein S3
		P63973	LigA, DNA ligase
P07814	EPRS, bifunctional aminoacyl-tRNA synthetase	P0A5U4	RecA, recombinase A
P53597	SUCLG1, succinyl-CoA ligase [ADP/GDP-forming] subunit alpha, mitochondrial	P71559	SucC, Succinyl-CoA ligase [ADP-forming] subunit beta
P25705	ATP5A1, ATP synthase subunit alpha, mitochondrial	P63677	AtpD, ATP synthase subunit alpha
		P0A620	TopA, DNA topoisomerase
		P63671	AtpG, ATP synthase gamma chain
P31153	MAT2A, S-adenosylmethionine synthase isoform type-2	P66875	MetB, Cystathionine gamma-synthase
		P0A528	ATP-dependent Clp protease ATP-binding subunit ClpX
		P63650	ScoB, probable succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B
		P0A644	RlmN, ribosomal RNA large subunit methyltransferase N
		P0A5U4	RecA, recombinase A
		P65801	Gcp, probable tRNA threonylcarbamoyladenosine biosynthesis protein Gcp
		P0A554	Dxs, 1-deoxy-D-xylulose-5-phosphate synthase
		P0A5Y8	SecA1, protein translocase subunit SecA1
		O53832	PstB1, phosphate import ATP-binding protein PstB 1
P47756	CAPZB, F-actin-capping protein subunit beta	P64411	HtpG, chaperone protein htpG
O15382	BCAT2, branched-chain-amino-acid aminotransferase, mitochondrial	P66875	MetB, cystathionine gamma-synthase
Q16740	CLPP, putative ATP-dependent Clp protease proteolytic subunit, mitochondrial	P0A520	GroL2, 60 kDa chaperonin 2
		P0A528	ATP-dependent Clp protease ATP-binding subunit ClpX
P12277	CKB, creatine kinase B-type	P64411	HtpG, chaperone protein htpG

		P71880	Mez, putative malate oxidoreductase [NAD]
P23921	RRM1, ribonucleoside-diphosphate reductase large subunit	O69664	GlpK, glycerol kinase
P00505	GOT2, aspartate aminotransferase, mitochondrial	P63937	MT0474, probable aldehyde dehydrogenase
P00558	PGK1, phosphoglycerate kinase 1	P96377	Eno, enolase
		P96901	Lhr, ATP-dependent helicase, putative
P49411	TUFM, elongation factor Tu, mitochondrial	P66537	RpsB, 30S ribosomal protein S2
		O69635	AcsA, acetyl-coenzyme A synthetase
		P0A520	GroL2, 60 kDa chaperonin 2
		P95052	RplB, 50S ribosomal protein L2
		P41194	RpsG, 30S ribosomal protein S7
P55809	OXCT1, succinyl-CoA:3-ketoacid-coenzyme A transferase 1, mitochondrial	P63650	ScoB, probable succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B
		P71715	DnaB, replicative DNA helicase
P61011	SRP54, signal recognition particle 54kDa protein	P77899	MetK, S-adenosylmethionine synthase
		P63673	AtpA, ATP synthase subunit alpha
P28340	POLD1, DNA polymerase delta catalytic subunit	P0A520	GroL2, 60 kDa chaperonin 2
P53582	METAP1, methionine aminopeptidase 1	P67510	LeuS, leucyl-tRNA synthetase
P13929	ENO3, beta-enolase	P71559	SucC, succinyl-CoA ligase [ADP-forming] subunit beta
		P65700	Pgk, phosphoglycerate kinase
		P96901	Lhr, ATP-dependent helicase, putative
		P69942	GlnE, glutamate-ammonia-ligase adenylyltransferase
O76031	CLPX, ATP-dependent Clp protease ATP-binding subunit clpX-like, mitochondrial	P77899	MetK, S-adenosylmethionine synthase
		P63288	ClpB, chaperone protein ClpB
		O53832	PstB1, phosphate import ATP-binding
		P0A5U4	RecA, recombinase A

		P0A526	ClpP1, ATP-dependent Clp protease proteolytic subunit 1
P04406	GAPDH, glyceraldehyde-3-phosphate dehydrogenase	P0A674	RpoC, DNA-directed RNA polymerase subunit beta'
		O06134	Pyk, pyruvate kinase
		P96901	Lhr, ATP-dependent helicase, putative
		P0A602	RpoD, RNA polymerase sigma factor RpoD
Q13951	CBFB, core-binding factor subunit beta	P0A558	Tuf, elongation factor Tu
P62244	RPS15A, 40S ribosomal protein S15a	P64411	HtpG, chaperone protein htpG
P11908	PRPS2, ribose-phosphate pyrophosphokinase 2	P95052	RplB, 50S ribosomal protein L2
		P67510	LeuS, leucyl-tRNA synthetase
		P0A520	GroL2, 60 kDa chaperonin 2
P06576	ATP5B, ATP synthase subunit beta, mitochondrial	P63673	AtpA, ATP synthase subunit alpha
		P63671	AtpG, ATP synthase gamma chain
Q9H1K1	ISCU, iron-sulfur cluster assembly enzyme ISCU, mitochondrial	P71558	SucD, succinyl-CoA ligase [ADP-forming] subunit alpha
P10768	ESD, S-formylglutathione hydrolase	O69664	GlpK, glycerol kinase
		P0A5Y8	SecA1, protein translocase subunit SecA 1
P54819	AK2, adenylate kinase 2, mitochondrial	P63852	CtaD, probable cytochrome c oxidase subunit 1
Q99704	DOK1, docking protein 1	P0A544	SerA, D-3-phosphoglycerate dehydrogenase
P49588	AARS, alanyl-tRNA synthetase, cytoplasmic	P0A636	GltX, glutamyl-tRNA synthetase
		P66016	PrfA, peptide chain release factor 1
		P0A620	TopA, DNA topoisomerase 1
P30405	PPIF, peptidyl-prolyl cis-trans isomerase F, mitochondrial	P96377	Eno, enolase
P11766	ADH5, alcohol dehydrogenase class-3	O53832	PstB1, phosphate import ATP-binding
Q6YP21	CCBL2, kynurenine-oxoglutarate transaminase 3	P63977	DnaE1, DNA polymerase III subunit alpha
P60174	TPI1, triosephosphate isomerase	P63977	DnaE1, DNA polymerase III subunit alpha

P0A644	RlmN, Ribosomal RNA large subunit methyltransferase N

Table 2.7: The "interolog-array" interactions

2.3.4.1 Connections between predicted interactions

Figure 2.5 represents the known and "interolog-array" interactions in the networks. Due to the size of the networks, especially the human network, only the first neighbours of the proteins involved in the inter-species interactions are shown. Figure 2.5 shows that there are 3 small modules formed by known *Mtb* proteins interacting with human proteins. The first module contains the *Mtb* protein PPE18 (Q7D8M9) which has 12 neighbour proteins. These proteins are members of the PE/PPE family and one protein is EsxK (O05299), which is an ESAT-6 like protein. The second module contains the proteins LpqH (P0A5J0) and Q50615. The neighbours of these two proteins are members of the PE/PGRS family, which is the largest class of the PE family. PE/PGRS proteins are characterized by a PE domain followed by a PGRS domain, which is a polymorphic GC-rich repetitive sequences [13]. The third module contains the alanine and proline-rich protein Apa (Q50906). This protein has as its neighbours putative uncharacterised proteins, uncharacterised PPE family protein, alanine and proline rich proteins and 3 proteins involved in the molybdenum transport system, which link this module to the second module. The "interolog-array" interactions are represented in Figure 2.6. This network has one large subnetwork containing 66 proteins. We compared the size of this largest connected component to that of randomly generated networks having the same characteristics as the "interolog-array" interactions. A random network is generated by randomly selecting 78 interactions among the 1,645 possible interactions between 35 human proteins and 47 *Mtb* proteins. Repeating this process 5,000 times gives a largest component of size 64, on average, and the probability of getting a largest connected component of size 66 is 0.1. The probability is given by counting how many times there are networks with largest component of size 66 during the 5,000 iterations and by dividing this number by the number of iterations.

Some of the human-*Mtb* protein-protein interactions are conserved interactions within one organism (shaded oval in Figure 2.6), and they include interactions between ATP synthase and Clp protease. Clp proteases are ATP-dependent proteases conserved among eubacteria and most eukaryotes. They are responsible for degrading abnormal proteins [5]. Clp proteases consist of two functional elements: a protease core (called ClpP and ClpQ) which can interact with different ATPase-active chaperones, namely ClpA, ClpC, ClpE and ClpX [71].

ATP synthase is a large membrane protein found in the plasma membrane of prokaryotes and in mitochondrial membranes of eukaryotes [27]. In most cases, they are responsible for ATP synthesis by creating a transmembrane proton gradient. They also work in the reverse direction: they hydrolyze ATP to generate a transmembrane proton gradient. ATP synthases consist of two structural domains: a membrane-bound F_O portion and a soluble F_1 portion. In *E. coli*, the F_1-ATPase consists of 5 subunits $\alpha, \beta, \gamma, \delta$ and ε and the F_O-ATPase consists of 3 subunits a, b and c [27]. In our interactions, ATP5A1 and ATP5B are the α and β subunits of the human F_1-ATPase, respectively, whereas atpA, atpD and atpG are the α, β and γ subunits of the *Mtb* F_1-ATP synthase, respectively.

2.3.5 Host-pathogen interaction prediction using inter-species interactions

This method is a variant of the interologs method presented in 2.3.2. Instead of using intra-species interactions to predict the interologs, we used experimentally verified interactions between human and bacterial proteins. From the inter-species human-pathogen interactions, if the pathogen has an ortholog in *Mtb*, we infer an interaction between human and *Mtb* protein. The experimentally verified host-pathogen interactions were downloaded from two databases dedicated to inter-species interactions, namely PATRIC and HPIDB (see 1.5.2). We filtered host-pathogen interactions from HPIDB to only use human-bacterial protein interactions. All the protein interactions from PATRIC were used since it contains only human-bacterial protein interactions. The ortholog data for *Mtb* are from INTEGR8. Figure 2.7 shows how human-*Mtb* interaction is inferred.

We predicted 845 and 195 human-*Mtb* interactions from PATRIC and HPIDB data, respectively. Combination of the two sets of interactions gave 1021 host-pathogen interactions between 784 human proteins and 216 *Mtb* proteins. Among these interactions, there were 866 where both proteins belong to the constructed networks (644 human and 204 from *Mtb*). These interactions were further investigated by considering only those differentially expressed during infection.

2.3.6 Filtering for "database-array" interactions

We filtered the 866 host-pathogen interactions previously found by taking only the interactions where both proteins are differentially expressed during infection. The same microarray experiment as in 2.3.4 was used to find differentially expressed genes. By applying the filter, we narrowed the interactions down to 109 interactions between 85 human proteins and

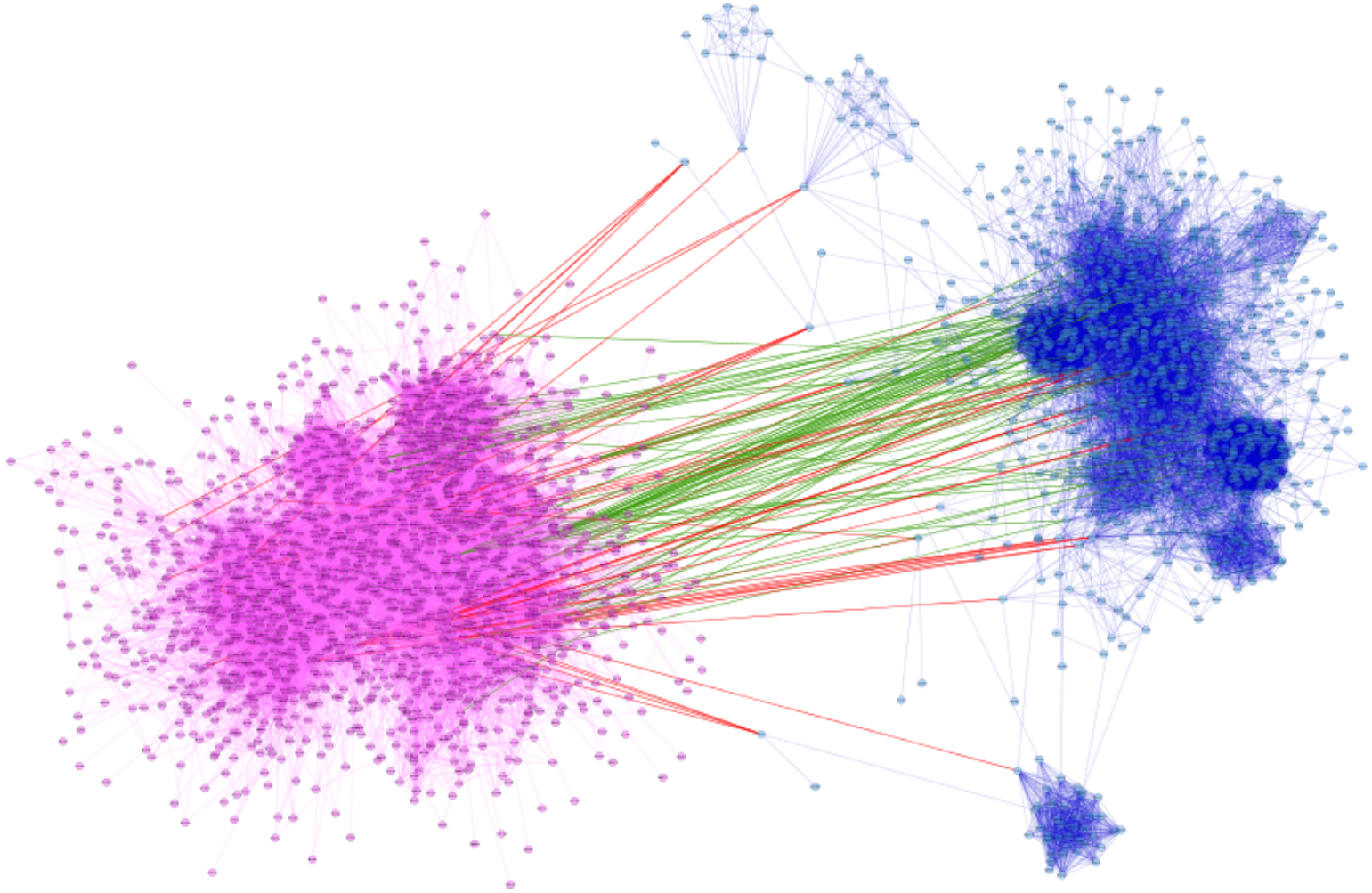


Figure 2.5: Known host-pathogen interactions with predicted host-pathogen interactions. Blue nodes represent pathogen proteins, pink nodes represent human proteins. Red lines are for known inter-species interactions, green lines are for "interolog-array" interactions, blue lines are for *Mtb* intra-species interactions and pink lines are for human intra-species interactions.

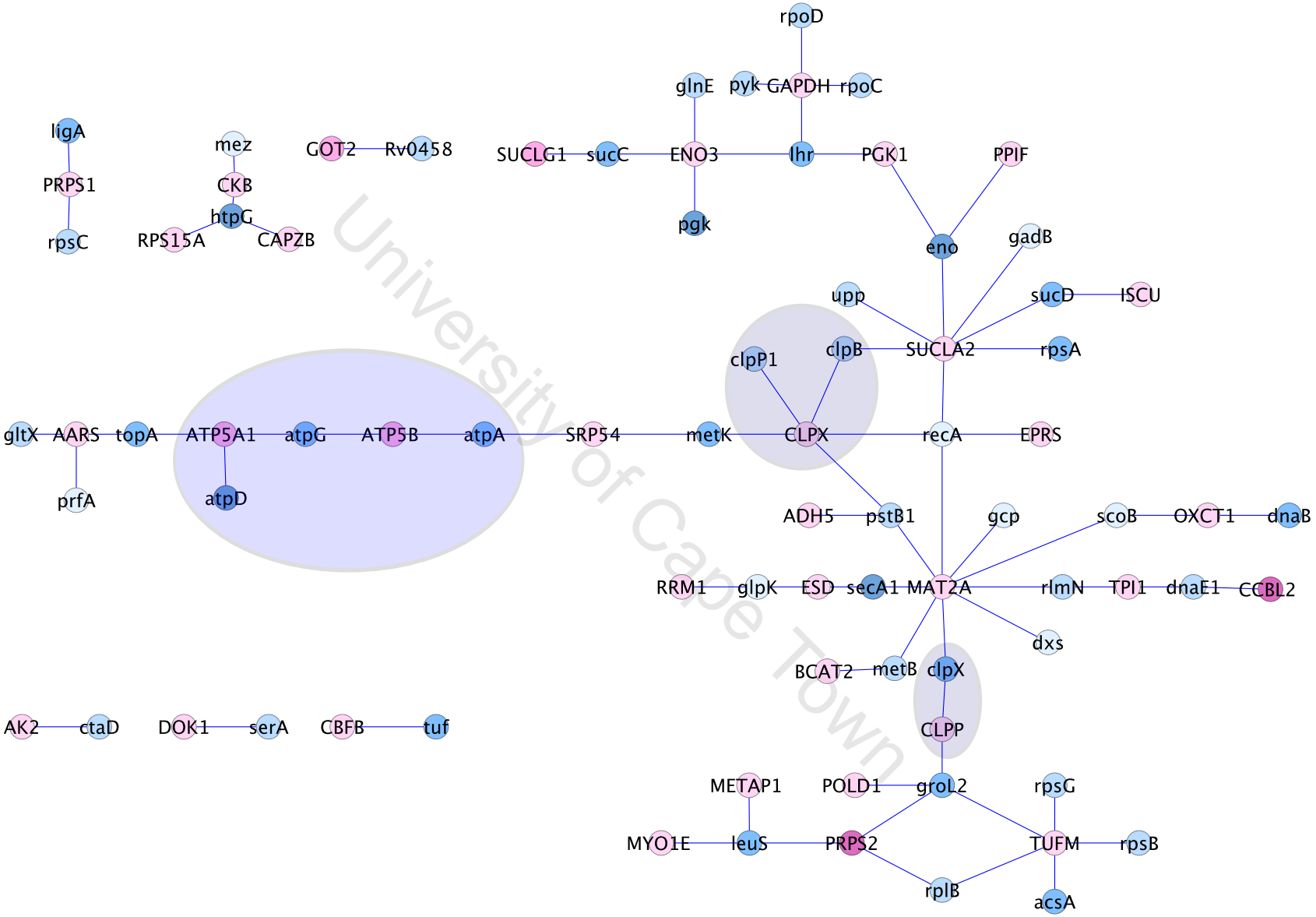


Figure 2.6: The "interolog-array" interactions.

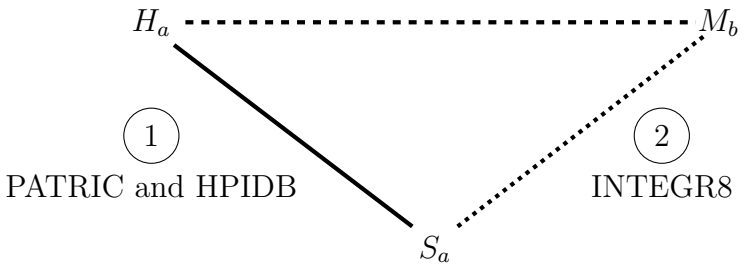


Figure 2.7: Human-*Mtb* interactions from inter-species interactions. The solid line represents experimentally verified host-pathogen interactions, the dotted line represents orthologs and the dashed line represents the inferred human-*Mtb* interaction.

53 *Mtb* proteins. We shall refer to the filtered interactions as "database-array" interactions. The initial interactions for the "database-array" interactions come from interactions between human and three bacterial species, namely *Bacillus anthracis*, *Francisella tularensis subsp. tularensis* and *Yersinia pestis*. This is due to the fact that a high-throughput yeast two-hybrid assay was used to identify physical interactions between human proteins and proteins from each of these three pathogens [44]. They therefore form most of the human-pathogen interactions from PATRIC and HPIDB.

2.3.6.1 Connections between predicted interactions

The "database-array" interactions are visualized in Figure 2.8 and with the first neighbour of the nodes in Figure 2.9. The network contains a large connected subnetwork of 61 proteins. As for the "interolog-array" interactions, we randomly generated networks by selecting 109 interactions among the possible interactions between 85 human proteins and 53 *Mtb* proteins and repeated the process 5,000 times. In this case, the average size of the largest connected component is 19 and the probability of getting a largest connected component of size 61 is 0.01.

In the largest subnetwork of the "database-array" interactions, we observe several proteins having 4 or more interactions. The human proteins NFKB1 and CD74 are connected to 10 and 5 *Mtb* proteins, respectively. Both of these proteins are known to play a role during tuberculosis infection. NFKB1 (nuclear factor NF-kappa-B subunit p105) is part of the NF-κB complex which controls the transcription of genes involved in the pro-inflammatory response as well as genes involved in the antiapoptotic response. During early infection, *Mtb* inhibits macrophage apoptosis by up-regulating the NF-κB signaling pathway, resulting in the up-regulation of FLIP, an inhibitor of death receptor signaling [80]. CD74 or HLA class II histocompatibility antigen gamma chain is implicated in the transport of MHC (major

histocompatibility complex) class II proteins from the endoplasmic reticulum to the Golgi complex. *Mtb* inhibits MHC class II antigen presentation which reduces the recognition of infected macrophages by CD_4^+T cells [59]. The *Mtb* proteins HemL (glutamate-1-semialdehyde aminotransferase), RpoD (RNA polymerase sigma factor rpoD), RecN (DNA repair protein recN), ThiC (thiamine biosynthesis protein ThiC), CydD (Transmembrane ATP-binding protein ABC transporter) and SdaA (L-serine dehydratase) are connected to 10, 4, 8, 5, 5 and 4 human proteins, respectively.

The "database-array" interaction network also has smaller subnetworks containing human proteins known to be relevant to tuberculosis infection. These proteins are JAK1, RAB5A, RAB5C, CTSD and CALM1. RAB5A (Ras-related protein Rab-5A), RAB5C (Ras-related protein Rab-5C) and CALM1 (Calmodulin) are all involved in the formation of the phagolysosome. CTSD (cathepsin D) is a protease secreted by the phagolysosome for bacterial degradation.

2.3.7 Connections between "interolog-array" and "database-array" interactions

In this section, we checked whether there are any common interactions shared by the two sets of host-pathogen interactions "interolog-array" and "database-array". There were no common interactions between the two networks. However, there were proteins shared by the two sets: 3 human proteins and 9 *Mtb* proteins (Figure 2.10). The shared human proteins are MAT2A (S-adenosylmethionine synthase isoform type-2), AK2 (Adenylate kinase 2, mitochondrial) and CAPZB (F-actin-capping protein subunit beta). The common *Mtb* proteins are DnaB (replicative DNA helicase), RlmN (Ribosomal RNA large subunit methyltransferase N), Upp (Uracil phosphoribosyltransferase), SecA1 (Protein translocase subunit SecA 1), RpoD (RNA polymerase sigma factor RpoD), TopA (DNA topoisomerase 1), GlnE (Glutamate-ammonia-ligase adenylyltransferase), LigA (DNA ligase) and SerA (D-3-phosphoglycerate dehydrogenase).

2.3.8 Discussion and conclusions

In this chapter, we constructed human and *Mtb* protein networks from various data. We retrieved 48 known host-pathogen interactions from the literature, and used the interolog method to predict host-pathogen interactions. 483 host-pathogen interactions were found by using intra-species interactions as a prediction method. We filtered the interologs using two criteria: 1) connections to known interactions (the "interolog-known" interactions); and

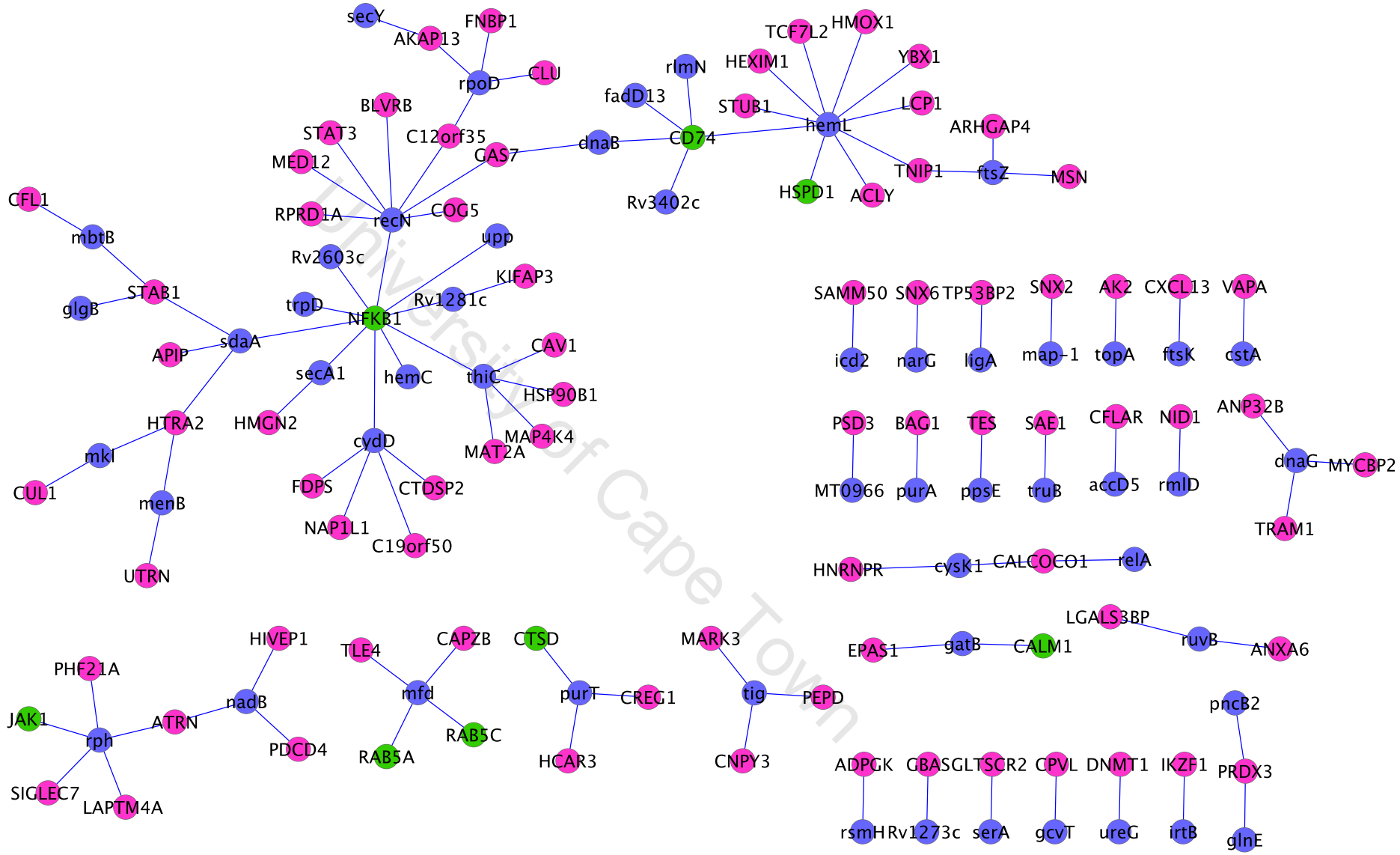


Figure 2.8: Network of "database-array" interactions. Pink and blue nodes represent *Mtb* and human proteins, respectively. Green nodes are human proteins known to be involved in tuberculosis pathogenesis.

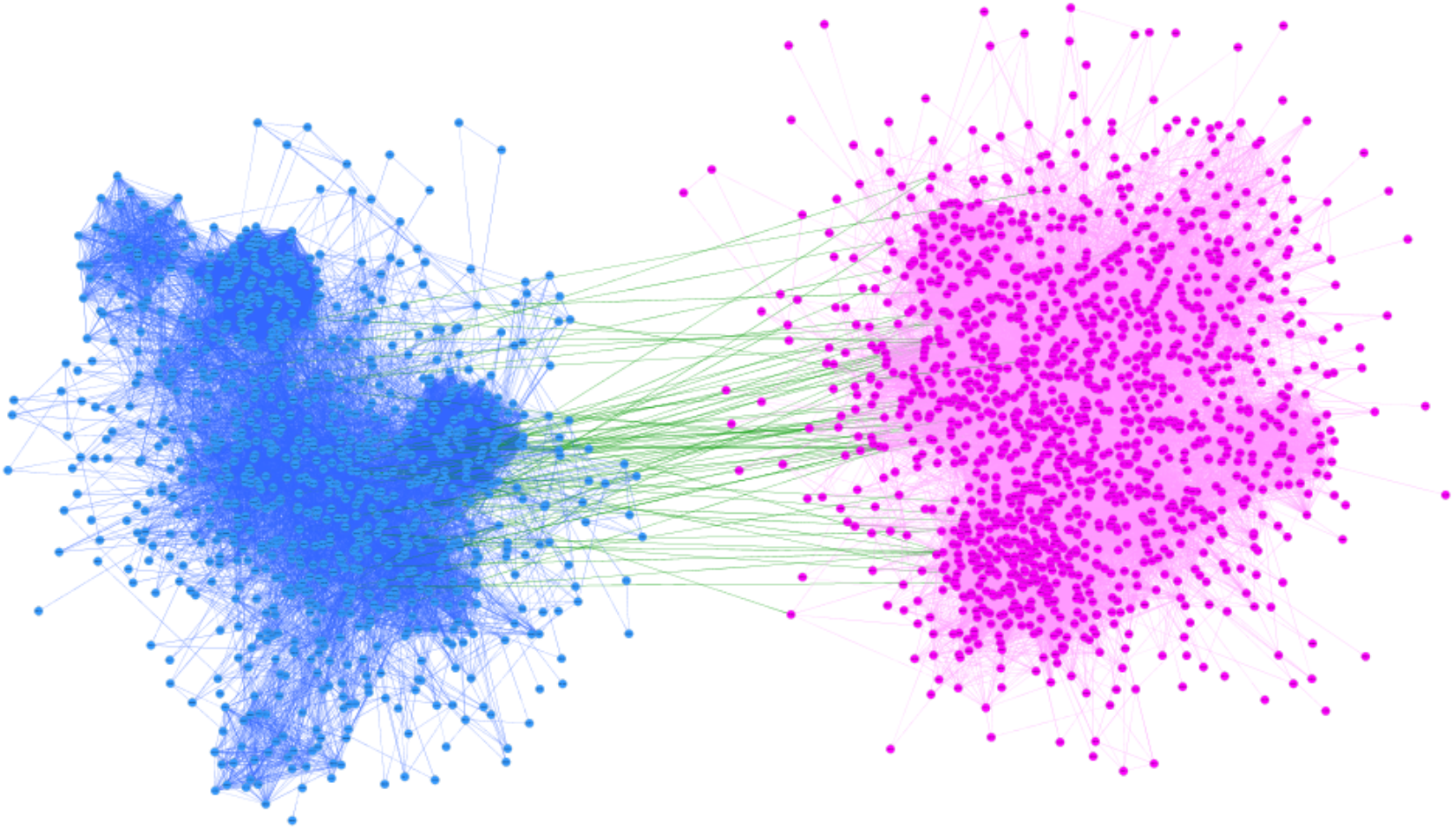


Figure 2.9: Network of "database-array" interactions with the first neighbours. Blue and pink nodes represent *Mtb* and human proteins, respectively.

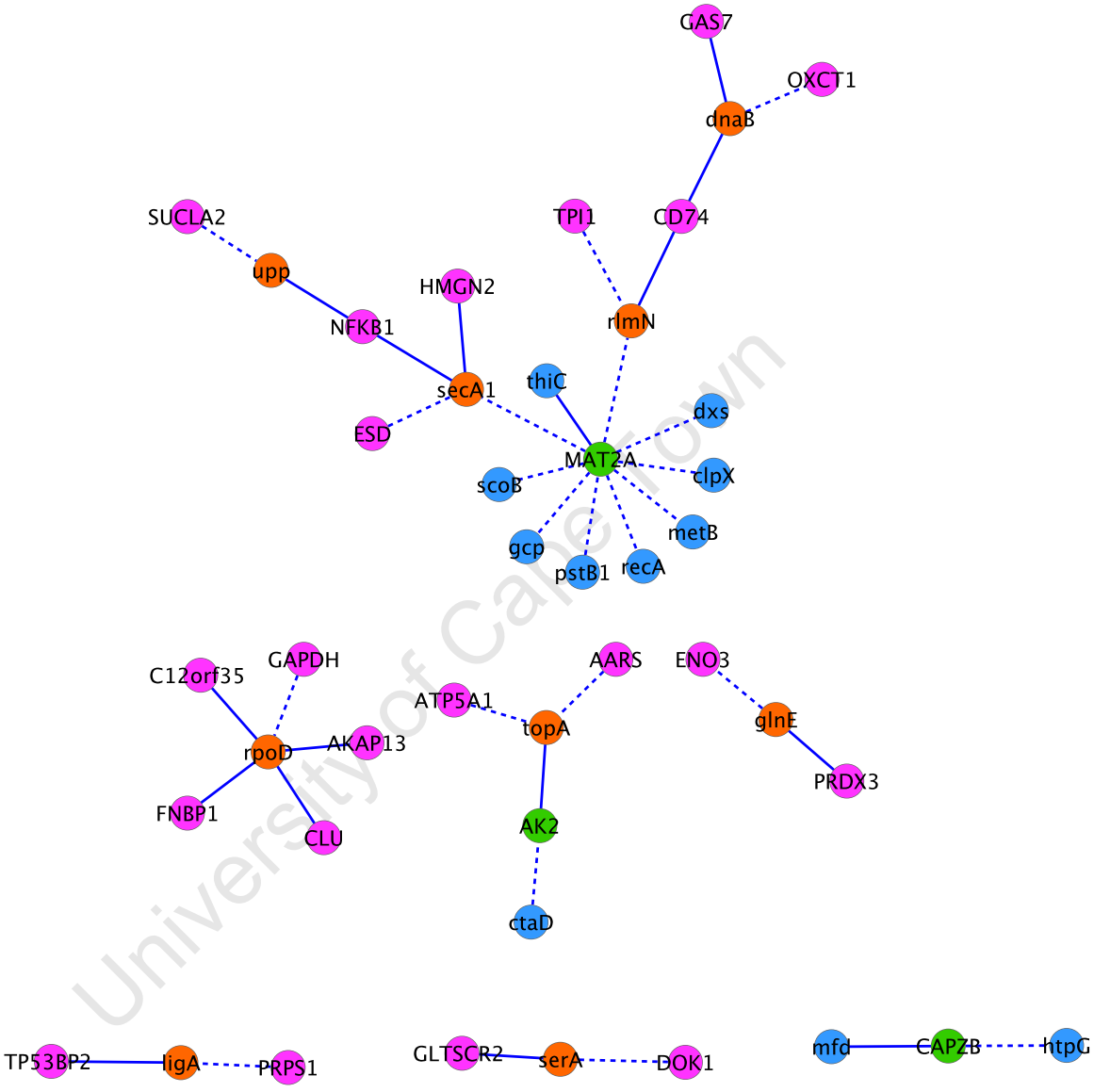


Figure 2.10: Common proteins between "interolog-array" and "database-array" interactions. Pink and blue nodes represent *Mtb* and human proteins, respectively. Orange nodes are the common *Mtb* proteins, and green nodes are the common common proteins. Dashed and solid lines are for "interolog-array" and "database-array" interactions, respectively.

2) whether they are expressed during infection (the "interolog-array" interactions) . The first filter gave 3 inter-species interactions whereas the second filter gave 78 inter-species interactions.

The interaction between EPRS (human) and RecA (*Mtb*) was predicted by both the "interolog-known" and "interolog-array", therefore increasing the confidence of the interaction. The inter-species interaction EPRS/RecA (P0A5U4) is the interolog of the intra-species interaction ProS/RecA (P78014) in *M. pneumoniae* M129. ProS is a proline-tRNA ligase. The interaction in *M. pneumoniae* M129 was experimentally verified by tandem-affinity purification [62].

We also used experimentally verified inter-species interactions from HPIDB and PATRIC to predict human-*Mtb* interactions. This approach yielded 866 inter-species interactions between 644 human and 204 *Mtb* proteins. These interactions were filtered based on their expression during tuberculosis infection. The resulting 109 interactions between 85 human proteins and 53 *Mtb* proteins were called the "database-array" interactions. There was no direct overlap between the "interolog-array" and "database-array" interactions. Nevertheless there were 3 human and 9 *Mtb* common proteins between the two sets of predicted interactions. We noted that there were no overlap between these two sets of interactions and the known host-pathogen interactions. A summary of the interactions in this chapter is presented in Table 2.8.

	Number of proteins	Number of interactions
Human network	16,548	334,070
Mtb network	4,070	38,049
Known interactions	14 human proteins 25 *Mtb* proteins	48
Interolog-known	3 human proteins 3*Mtb* proteins	3
Interolog-array	35 human proteins 47 *Mtb* proteins	78
Database-array	85 human proteins 53 *Mtb* proteins	109

Table 2.8: Summary of the constructed networks and predicted interactions.

Though we used known interactions and microarray data to narrow down the interactions, it was important to put the interactions in their biological context to see if they are likely to happen, which was the aim of the next chapter.

Chapter 3

Analysis of predicted host-pathogen interactions

In the previous chapter, we constructed a human and *Mtb* protein network and predicted inter-species interactions between the two organisms using the interolog method. This yielded a total of 483 host-pathogen interactions involving 175 human and 192 *Mtb* proteins. We applied two filters to find interactions relevant to tuberculosis infection. The first filter consists of the interologs where both proteins are neighbours of known proteins that we called "interolog-known" interactions. The second filter consists of the interologs where both proteins are differentially expressed during infection according to the microarray data from Bossi and Lehner. We called these interactions "interolog-array". The two filters allowed us to retrieve 3 and 78 interactions, respectively. We also predicted inter-species interactions using experimentally verified host-pathogen interactions that we called "database-array" interactions. We retrieved 109 such interactions. In this chapter, we shall focus on the "interolog-known", "interolog-array" and "database-array" interactions by analysing some features of the predicted interactions, including their subcellular locations, GO biological processes and their pathways.

3.1 The Gene Ontology

With the large increase in genomic data, there is a great need to represent and process information about genes, their products and their functions across all species in a standardized way. For this reason, the Gene Ontology (GO) consortium has developed a set of structured, controlled vocabularies for the annotation of genes and gene products [6].

GO is divided into three ontologies describing molecular functions (MF), biological processes (BP) and cellular components (CC).

Molecular function describes the function of a gene product at a molecular level. The functions of a gene product are the jobs that it performs and may include transporting things around the cell, binding, holding things together and converting one molecule into another. *Catalytic activity* and *protein binding* are examples of molecular function.

A biological process is defined in GO as a "series of events accomplished by one or more ordered assemblies of molecular functions". A process has a defined beginning and end. Biological process terms can be quite specific (*pyrimidine metabolic process* or *alpha-glucoside transport*) or very general (*signal transduction* or *cellular physiological process*). A biological process is generally distinguished from molecular function by the fact that processes must have more than one distinct step.

Finally, a cellular component describes locations of the gene product, which can be subcellular structures (*nucleus* or *nuclear inner membrane*) or macromolecular complexes (*ubiquitin ligase complex* or *ribosome*). Extracellular environments of cells are also included in the cellular component ontology.

Ontologies in GO are structured as directed acyclic graphs (DAGs) in which terms represent the nodes and arcs represent the relations between the terms. Therefore, terms have parent-child relationships, with child terms being more specific or specialized than their respective parent(s). For example, in the cellular component ontology, centrosome is a child of intracellular non-membrane bounded organelle, which is in turn a child of organelle. The relations between GO terms comprise *is a*; *part of*; and *regulates*, *negatively regulates* and *positively regulates* (http://www.geneontology.org/GO.ontology.relations.shtml).

The *is_a* relation in GO means subtype. For example, if we say mitotic cell cycle *is a* cell cycle, it means mitotic cell cycle is a subtype of cell cycle. The *is_a* relation is transitive, that is to say if A *is_a* B and B *is_a* C, then A *is_a* C.

The *part_of* relation is described as follows: if B is *part_of* A, then if B exists, A necessarily exists and B is part of A. However, the existence of A does not necessarily imply the existence of B. For example replication fork is *part of* chromosome, means all replication forks are part of some chromosome, but only some chromosomes have a part replication fork. Like the *is a* relation, the *part_of* relation behaves transitively.

The *regulates*, *positively regulates* and *negatively regulates* relations denote biological regulation.

GO has numerous applications such as:

- integrating protein information from different organsims
- assigning functions to protein domains
- finding functional similarities in genes that are coexpressed under certain conditions
- predicting the likelihood that a particular gene is involved in diseases that haven't yet been mapped to specific genes
- developing automated ways of deriving information about gene function from the literature
- verifying models of genetic, metabolic and product interaction networks.

AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi), developed by the GO consortium, is the official web-based toolset for searching and browsing the Gene Ontology database [29]. QuickGO, developed by the European Bioinformatics Institute (EBI), also provides web-based access to the GO database [16]. However, while AmiGO and QuickGO are good for accessing small datasets, the Gene Ontology Annotation (Uniprot-GOA) Database [25], also developed by EBI, is better fitted to large data and provides high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). The GOA database can be accessed at http://www.ebi.ac.uk/GOA/.

We used GO to assess the functions of interacting proteins and to determine whether it is biologically possible for our predicted interactions to take place. For example, within an organism if two proteins share the same location (using the GO cellular component), they are likely to interact, and across organisms, the interacting proteins must have the potential to encounter each other.

3.2 Cellular component analysis

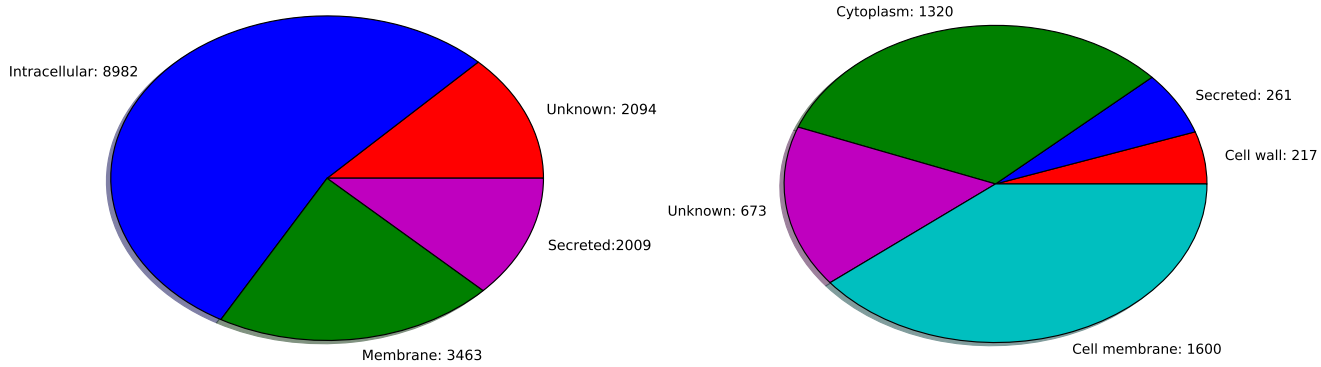
Protein interactions are constrained by physical location, meaning that two proteins can interact only if they have the appropriate subcellular location.

In order to find the location of the human and *Mtb* proteins, we used their GO cellular component terms. We converted these to GO slim terms associated with the cellular component instead of the full GO cellular component ontology. GO slims are truncated versions of the GO ontologies where the more specific terms are mapped to one or more high-level terms. GO slims may be user defined or predefined by the GO Consortium GO slims. The "Investigate GO slims option" of the QuickGO tool from EBI (http://www.ebi.ac.uk/QuickGO/)

was used to find the GO slim cellular component terms of the proteins. The tool has three main parts: "Choose Terms", "Refine Selection" and "Find Annotations". In the "Choose Terms", we choose the GO slim set we want to use. We can directly select one of the predefined GO slims created by groups in the GO consortium. The "Add terms" text box allows us to create a customised version of the predefined GO slim sets or create a new GO slim set by adding GO identifiers in the text box. In the "Refine Selection", we can remove a GO term from the set we have chosen or remove all terms in a particular category (BP, MF or CC). In our case, we removed the biolocal process and molecular function category and only kept the cellular component category. Finally, the "Find Annotations" displays the annotations of the proteins related to our GO slim set. However, it displays all the proteins in UniProt. Unwanted proteins can be filtered out by choosing the "Filter" feature. Several parameters, such as sequence identifier, taxon, or GO ID can be used as filters.

For the human proteins, we used the GO slim Generic, which is a predefined GO slim set. The set is not species specific but it has many terms pertaining to eukaryotic organisms. The set contains 35 cellular components terms. We used all the protein IDs of the human network to filter the annotation. We then downloaded the results and collected the GO cellular components that we divided into three groups: intracellular, membrane and secreted. We also have the unknown category containing proteins that have no cellular component annotation. The intracellular group contains the terms referring to all components inside of the cells. The membrane group only contains the plasma membrane and the secreted group contains the proteins in the extracellular region. Since a protein may be annotated to more than one group, we set the final annotation of the protein as the furthest it can go outside the cell. The repartition of the human protein locations is presented in Figure 3.1a. The intracellular proteins form about half of the proteins, around 21% of the proteins are found in the membrane, 12.14% are secreted and 12.65% of the proteins have unknown locations.

We used only the GO slim terms relevant to prokaryotes in the GO slim generic to build the GO slim terms for the *Mtb* proteins and we used all the protein IDs of the *Mtb* network to filter the annotation. In the case where the protein had no subcellular location, we used PSORTb to predict its location, which was the case for 2272 proteins. PSORTb is a web-based tool for prediction of subcellular location in bacterial protein sequences [150]. To predict subcellular location using PSORTb, we first downloaded the FASTA sequences of the proteins from UniProt. The FASTA file was then used as input to PSORTb. The parameters of PSORTb were set as follows: the organism type was set to bacteria and the Gram stain was set to positive. The result was a tab-delimited file containing the sequence ID of the protein, the final prediction of localization and the score, each line representing one protein. The final prediction of localization of proteins were "Cytoplasmic", "CytoplasmicMembrane",



(a) Human (b) *Mtb*

Figure 3.1: Repartition of human and *Mtb* proteins in the different localizations

"Cellwall", "Extracellular" and "Unknown". We used these locations to divide the predicted locations into 5 main categories, from the inside of the cell to the outside: cytoplasm, cell membrane, cell wall, secreted and unknown. Secreted proteins are proteins secreted into the cell surroundings and proteins whose locations could not be identified were classed as "unknown". Again, if a protein was annotated by GO with more than one location, then we took the one that is furthest from the inside of the cell. The repartition of locations for *Mtb* is shown in Figure 3.1b.

3.2.1 *Mtb* protein subcellular locations

M. tuberculosis is a prokaryotic cell, it does not have a nucleus or membrane-bound organelles. An *Mtb* cell consists of:

- a peptidoglycan cell wall enclosing a cytoplasmic membrane or cell membrane, which consists of a lipid bilayer with embedded proteins,
- a cytoplasm containing proteins and genetic material,
- external structures such as flagella and pili.

Table 3.1 shows the number of proteins in each subcellular location for the *Mtb* proteins that are involved in the known, "interolog-known", "interolog-array" and "database-array"

interactions. *Mtb* proteins located in the cytoplasm are not likely to interact with host proteins unless the Mtb cells are lysed, whereas proteins located in the cell membrane and those which are secreted are more likely to interact with host proteins.

Subcellular location	Known	Interolog-known	Interolog-array	Database-array
Cytoplasm	0	1	8	9
Cell membrane	4	1	19	28
Cell wall	5	1	15	8
Secreted	16	0	5	8
Unknown	0	0	0	0

Table 3.1: Number of proteins in the 5 subcellular locations for *Mtb* proteins involved in the known, "interolog-known", "interolog-array" and "database-array" interactions

3.2.2 Human protein subcellular location

Unlike *Mtb*, human cells are eukaryotic cells, which have more complex inner structures than prokaryotic cells. They contain membrane-bound compartments, called organelles, which serve specific functions such as energy production and protein synthesis. The nucleus is perhaps the most important among the eukaryotic organelles because it is the location of a cell's DNA. Figure 3.2 represents an animal eukaryotic cell with its various organelles and a summary of the locations of interacting human proteins is provided in Table 3.2.

Subcellular location	Known	Interolog-known	Interolog-array	Database-array
Intracellular	1	2	29	45
Membrane	6	1	4	23
Secreted	7	0	0	12
Unknown	0	0	2	5

Table 3.2: Number of interacting human proteins in the intracellular, membrane, secreted and unknown categories for proteins involved in the known, "interolog-known", "interolog-array" and "database-array" interactions

As seen in Table 3.2, all but one of the human proteins involved in known interactions are located in the membrane or are secreted. The known interactions are generally derived from the well studied process of recognition and uptake of *Mtb* by the macrophages and the

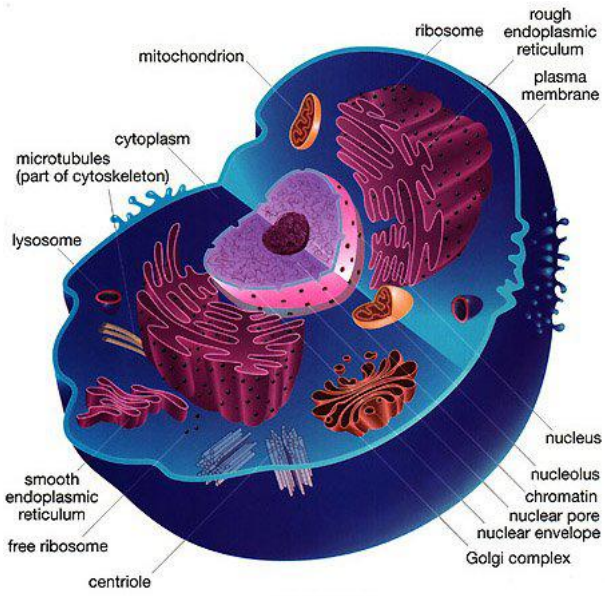


Figure 3.2: Organelles of an animal eukaryotic cell (http://www.uvm.edu/~inquiryb/webquest/fa06/mvogenbe/Animal-Cell.jpg).

dendritic cells. On the other hand, almost all the human proteins predicted to interact with *Mtb* proteins are located within the cell. This could be explained by the fact that these proteins are expressed during infection of macrophages and dendritic cells by *Mtb*, therefore the mycobacteria is located within the cell and is likely to interact with proteins that are intracellular.

3.2.3 Host-pathogen protein subcellular location

After looking at the protein locations from the host and the pathogen individually, we analysed the locations of the host-pathogen protein pairs. Table 3.3 shows the localization of the host-pathogen protein pairs.

As seen in Table 3.3, for the known host-pathogen protein interactions, the human proteins are located in the membrane or secreted; and the *Mtb* proteins are located in the cell membrane, cell wall or secreted. This is in agreement with the studies the interactions come from, which focus on the recognition of *Mtb* by human receptors which are located on the cell surface. The *Mtb* proteins are also located on the cell surface to enable binding to the human receptors. One exception is the interaction between the human heat shock protein 40 (HSP40) and the *Mtb* antigen mpt64. The human protein is annotated as intracellular and the bacterial protein is annotated as secreted, which is in accordance with [32], co-localizing mpt64 protein and HSP40 in the cytoplasm of HeLa cells.

Mtb is an intracellular pathogen that has evolved strategies to survive in intracellular phagosomes. A study by van der Wel *et al.* shows that after two days of infection, *Mtb* progressively translocates from phagolysosomes into the cytosol in nonapoptotic cells [135]. Therefore, the likely locations of the host-pathogen protein pairs are intracellular-cell-wall/cell-membrane/secreted if the pathogen is within the host cell, or secreted-cell-wall/cell-membrane/-secreted if the pathogen just encounters the host cell.

In addition, although GO and PSORTb predicted some of the interacting *Mtb* proteins to be cytoplasmic, annotation for some, ScoB, Glpk and GadB, for example, suggests that these proteins have been identified by mass spectrometry in the membrane fraction of *Mtb* cells. Therefore the subcellular location analysis provides us with more biological insight into the predicted interactions but we did not want to rule out possible interactions based on subcellular locations.

Host protein location	Mtb protein location	Known	Interolog-known	Interolog-array	Database-array
Intracellular	Secreted	1	0	9	8
Intracellular	Cytoplasm	0	1	13	5
Intracellular	Cell membrane	0	1	23	35
Intracellular	Cell wall	0	0	22	11
Membrane	Cell membrane	5	0	1	23
Membrane	Cytoplasm	0	0	0	3
Membrane	Cell wall	2	1	5	4
Membrane	Secreted	10	0	1	1
Secreted	Cytoplasm	0	0	0	5
Secreted	Cell membrane	6	0	0	7
Secreted	Cell wall	3	0	0	0
Secreted	Secreted	21	0	0	1
Unknown	Cytoplasm	0	0	0	1
Unknown	Cell membrane	0	0	2	3
Unknown	Cell wall	0	0	2	2

Table 3.3: Number of proteins in each host-pathogen location pair for proteins involved in the known,"interolog-known", "interolog-array" and "database-array" interactions

3.3 Gene ontology term enrichment

To gain insight into the biological processes of the proteins involved in host-pathogen interactions, we perform a GO term enrichment analysis using the DAVID (Database for Annotation, Visualization and Integrated Discovery) Functional Annotation Chart tool (http://david.abcc.ncifcrf.gov/). A GO term enrichment analysis is used to highlight the most relevant GO terms associated with a given gene list compared to the genes from which the list is derived, known as the background. DAVID uses a modified Fisher's exact p-value, called EASE score, to calculate the enrichment p-value [63]. In order for DAVID to compute a GO term enrichment analysis, it needs a gene list for which we want to see the enrichment and a background list. Next, the annotation categories of interest have to be selected. In our case, we chose the Gene Ontology category, more particularly the GO biological process category as we want to perform a GO biological process term enrichment. For each protein list of interest ("interolog-array" or "database-array" proteins), we performed a GO biological process enrichment analysis using the human proteins or *Mtb* proteins from our constructed network as background. As suggested on the DAVID website, we selected the GO terms situated in the fourth and fifth level to avoid very general GO terms such as "biological process". We used the Bonferonni p-value, which is a corrected p-values for multiple testing, and we selected those GO terms enriched in our candidate protein list by requiring a p-value less than 0.05.

The human "interolog-array" proteins predicted to interact with *Mtb* proteins were enriched in nitrogen compound biosynthetic process, oxoacid metabolic process, carboxylic acid metabolic process and nucleobase, nucleoside and nucleotide metabolic process. When *Mtb* cells are phagocytozed they are exposed to nitric oxide and oxidative stress. The interacting human proteins may be involved in facilitating this. On the other hand, the human proteins predicted using databases are enriched in processes related to negative regulation of apoptosis and positive regulation of cellular process (Table 3.4). It is known that virulent *Mtb* strains inhibit apoptosis of the host macrophage to protect their replicative niche [76].

For the *Mtb* proteins, the smallest Bonferonni corrected p-values were 0.100316 and 0.345477 for the "interolog-array" and "database-array", respectively, meaning that we have insufficient evidence to support the claim that our gene list is enriched in any GO biological processes. Nevertheless, when looking at the individual proteins, some of them share the same biological processes with *Mtb* proteins known to interact with human. These proteins and their GO terms are presented in Table 3.5. The biological processes include those particularly relevant to the intracellular environment of *Mtb* in the host, such as growth of symbiont in host, response to stresses related to the harsh intracellular environment, response to host

GO Term	Annotation	No. of genes	p-value	Bonferonni Test
GO:0044271	Nitrogen compound biosynthetic process	9	1.04×10^{-6}	0.000238
GO:0043436	Oxoacid metabolic process	9	5.42×10^{-5}	0.012279
GO:0019752	Carboxylic acid metabolic process	9	6.52×10^{-5}	0.014697
GO:0055086	Nucleobase, nucleoside and nucleotide metabolic process	7	1.19×10^{-4}	0.026762
GO:0048522	Positive regulation of cellular process	27	2.03×10^{-5}	0.012038
GO:0042981	Regulation of apoptosis	17	2.19×10^{-5}	0.014412
GO:0043067	Regulation of programmed cell death	17	2.47×10^{-5}	0.01462
GO:0010941	Regulation of cell death	17	2.59×10^{-5}	0.015313
GO:0043066	Negative regulation of apoptosis	11	6.10×10^{-5}	0.035683
GO:0043069	Negative regulation of programmed cell death	11	6.88×10^{-5}	0.040119
GO:0060548	Negative regulation of cell death	11	7.04×10^{-5}	0.041059

Table 3.4: Enriched GO biological process terms in human "interolog-array" proteins (top) and human "database-array" proteins (bottom).

immune response, and pathogenesis. Proteins playing a role in these processes may achieve their goal in protecting the pathogen from the environment through interaction with host proteins.

Furthermore, 32 and 22 proteins from the "interolog-array" and the "database-array" *Mtb* lists, respectively, were among the essential genes required for growth identified by Sassetti *et al.* using transposon site hybridization (TraSH) [116]. The genes in Sassetti *et al.* were initially for *Mtb* H37Rv but the orthologue file from INTEGR8 allowed us to map them to CDC1551 genes. In particular, 3 "database-array" proteins are also virulence factors of *Mtb*, namely NarG (O06559), RelA (P66014) and MbtB (P71717). NarG is a nitrate reductase. Nitrate respiration helps the bacteria to survive in O_2-depleted areas of inflammatory or necrotic tissue. NarG interacts with the human protein sorting nexin SNX6 (Q9UNH7), which is thought to be involved in several stages of intracellular trafficking. RelA is a protein that coordinates the metabolism of (p)ppGpp, a mixture of 3'-pyrophosphate derivative of GDP (ppGpp) and 3'-pyrophosphate derivative of GTP (pppGpp). Under stress condi-

UniProt Acc	Protein name	GO ID	GO name
O53832	PstB1	GO:0035435	phosphate ion transmembrane transport
		GO:0044117	growth of symbiont in host
P0A558	Tuf	GO:0001666	response to hypoxia
		GO:0006184	GTP catabolic process
		GO:00100039	response to iron ion
P0A602	RpoD	GO:0009405	pathogenesis
		GO:0009415	response to water
		GO:0052572	response to host immune response
P63288	ClpB	GO:0006950	response to stress
		GO:0009408	response to heat
P63650	ScoB	GO:0001666	response to hypoxia
P63852	CtaD	GO:0009060	aerobic respiration
P64411	HtpG	GO:0006950	response to stress
		GO:0071451	cellular response to superoxide
P69942	GlnE	GO:0040007	response to ammonium ion
O53306	FadD13	GO:0001101	response to acid
		GO:0044119	growth of symbiont in host cell
P63393	IrtB	GO:0009405	pathogenesis
O53189	Tig	GO:0006457	protein folding
		GO:0009267	cellular response to starvation
		GO:0046677	response to antibiotic
O06559	NarG	GO:0001101	response to acid
		GO:0001666	response to hypoxia
P95095	CstA	GO:0009267	cellular response to starvation
P66014	RelA	GO:0009405	pathogenesis
P71717	MbtB	GO:0009405	pathogenesis
		GO:0052572	response to host immune response
Q50723	Rv3402c	GO:0052572	response to host immune response
P0A602	RpoD	GO:0009405	pathogenesis
		GO:0052572	response to host immune response

Table 3.5: Important GO biological processes of *Mtb* "interolog-array" proteins (top) and "database-array" proteins (bottom)

tions, such as nutrient starvation, RelA produces (p)ppGpp that accumulates intracellularly and suppresses synthesis of stable RNA, induces degradative pathways and modulates ex-

pression of genes involved in DNA replication [105]. RelA was predicted to interact with the human protein CALCOCO1 (Q9P1Z2). CALCOCO1 is thought to be involved in elementary cellular functions linked to Ca^{2+}/calmodulin signaling [129]. Finally, MbtB, also known as mycobactin synthetase protein B, is involved in the initial steps of the mycobactin biosynthetic pathway. Mycobactins are lipophilic siderophores of mycobacteria mediating iron acquisition within macrophages [82]. MbtB interacts with the human proteins stabilin-1 STAB1 (Q9NY15) and cofilin-1 CFL1 (P23528). Stabilin-1 is a scavenger receptor and binds to both Gram-positive and Gram-negative bacteria [1] whereas cofilin-1 is a 18kDa phosphoprotein that regulates actin cytoskeleton dynamics.

3.4 GO biological process similarity score

We also functionally compared predicted human-*Mtb* protein pairs by computing their GO biological process similarity. Semantic similarity measures allow us to compare similarities between proteins based on their annotations [103]. We used the GO-universal metric based on the GO-DAG topology to compute the biological similarity between two proteins [86]. For each pair of proteins involved in a predicted host-pathogen interaction, we computed the similarity between their GO annotations. For each protein participating in the host-pathogen interactions, we also computed the similarity between that protein and all its neighbours. Table 3.6 represents the similarity score for each host-pathogen interaction list.

On average, the inter-species interactions have a low functional similarity, particularly compared to the similarity scores between intra-species neighbours. For the known host-pathogen interactions, two interactions have a similarity score of 0. The interactions are between the *Mtb* antigen 85-A and the human proteins fibronectin and plasminogen. In the "database-array" interactions, 7 host-pathogen interactions had a similarity score of 0. These interactions are listed in Table 3.7.

However, two inter-species interactions have a high functional similarity as is the case of the "interolog-array" interaction between AARS (P49588) and GltX (P0A636), which has the maximum similarity score among the "interolog-array" interactions. AARS is a cytoplasmic alanine–tRNA ligase whereas GltX is a glutamate–tRNA ligase. The interactions between TLE4 (Q04727), a transducin-like enhancer protein 4, and Mfd (P64326), a transcription-repair-coupling factor, has the highest similarity score among the "database-array" interactions. Apart from these cases, it is perhaps not surprising that inter-species interactions would have lower similiarity scores, as these interactions are more likely to be between proteins from different pathways than for intra-species interactions.

	Human similarity score with neighbour	*Mtb* similarity score with neighbour	Inter-species similarity score
Minimum	0.0957	0.00104	0.0
Mean	0.2781	0.24062	0.10190
Maximum	0.4389	0.75508	0.33847
Minimum	0.3129	0.2944	0.06425
Mean	0.3527	0.3885	0.11012
Maximum	0.3934	0.5704	0.13411
Minimum	0.08021	0.1545	0.00638
Mean	0.31987	0.3743	0.24065
Maximum	0.61023	0.6846	0.76297
Minimum	0.03812	0.003152	0.0
Mean	0.30295	0.334651	0.08372
Maximum	0.74204	0.65133	0.62843

Table 3.6: Similarity scores for the various lists of host-pathogen interactions. The first row is for the known interactions, the second row for the "interolog-known", the third row for the "interolog-array" and the last row for the "database-array" interactions.

3.5 Pathways

Bacterial pathogens attack intracellular-signalling and cytoskeletal pathways to alter host responses in a way that benefits them. For example, *Mtb* interferes with the NF-κB and the MAPK signalling pathways leading to the prevention of NF-κB dependent transcription and the alteration of antigen presentation, respectively (see 1.3.3). Bacteria might also hijack host metabolic pathways. For instance, *Chlamydia pistacci* hijacks the host's tryptophan depletion pathway by intercepting the byproduct kynurenine, which is used by *C. pistacci* to produce its own tryptophan [49].

We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) to find pathways that the predicted proteins belong to. The KEGG Mapper-Search and Color Pathway mapping tool allows us to search given objects, such as genes, proteins or compounds against KEGG pathways maps. For each list of predicted proteins (human and *Mtb*, "interolog-array" or "database-array"), we obtained a list of pathways relevant to the protein list. However, not all proteins belong to a pathway and a protein may be part of several pathways. For each human-*Mtb* protein interaction with both proteins belonging to one or more pathways, we looked for common pathways, substrates or products, as well as pathways involved in

Human UniProt Acc	Human protein name	*Mtb* UniProt Acc	*Mtb* protein name
P08133	ANXA6, annexin A6	P66753	RuvB, Holliday junction ATP-dependent DNA helicase
Q9NYI0	PSD3, PH and SEC7 domain-containing protein 3	O86346	MT0966, fumarylacetoacetate hydrolase family/metallo-beta-lactamase superfamily protein
O75882	ATRN, attractin	P65499	NadB, L-aspartate oxidase
O75882	ATRN, attractin	Q10628	Rph, ribonuclease PH
P46939	UTRN, utrophin	O06414	MenB, DHNA-CoA synthase
O75629	CREG1, Cellular repressor of E1A-stimulated genes 1	P95197	PurT, Phosphoribosylglycinamide formyltransferase 2
Q15012	LAPTM4A, lysosomal-associated transmembrane protein 4	Q10628	Rph, ribonuclease PH

Table 3.7: The "database-array" interactions having a similarity score of 0.

tuberculosis infection.

By applying this method, we were able to reconstruct a small network centered on the human MAT2A protein (Figure 3.3). MAT2A catalyzes the reaction from L-methionine to S-adenosyl-L-methionine in the cysteine and methionine metabolism. MAT2A was predicted to interact with 10 *Mtb* proteins in total. 3 of these proteins, namely MetK, RlmN and ThiC, have direct links to L-methionine or S-adenosyl-L-methionine in metabolic pathways. MetK is a methionine adenosyltransferase and performs the same function as MAT2A. RlmN is a ribosomal RNA large subunit methyltransferase N and specifically methylates position 2 of adenine 2503 in 23S rRNA. Finally, ThiC is a hydroxymethylpyrimidine phosphate synthase catalyzing the synthesis of the hydroxymethylpyrimidine phosphate (HMP-P) moiety of thiamine from aminoimidazole ribotide (AIR) in a radical S-adenosyl-L-methionine dependent reaction. ThiC is an essential gene necessary for thiamin (vitamin B1) synthesis [115]. Two proteins interacting with MAT2A, namely Dxs and ScoB, belong to pathways using S-adenosyl-L-methionine. Dxs catalyses the formation of 1-deoxy-D-xylulose 5-phosphate,

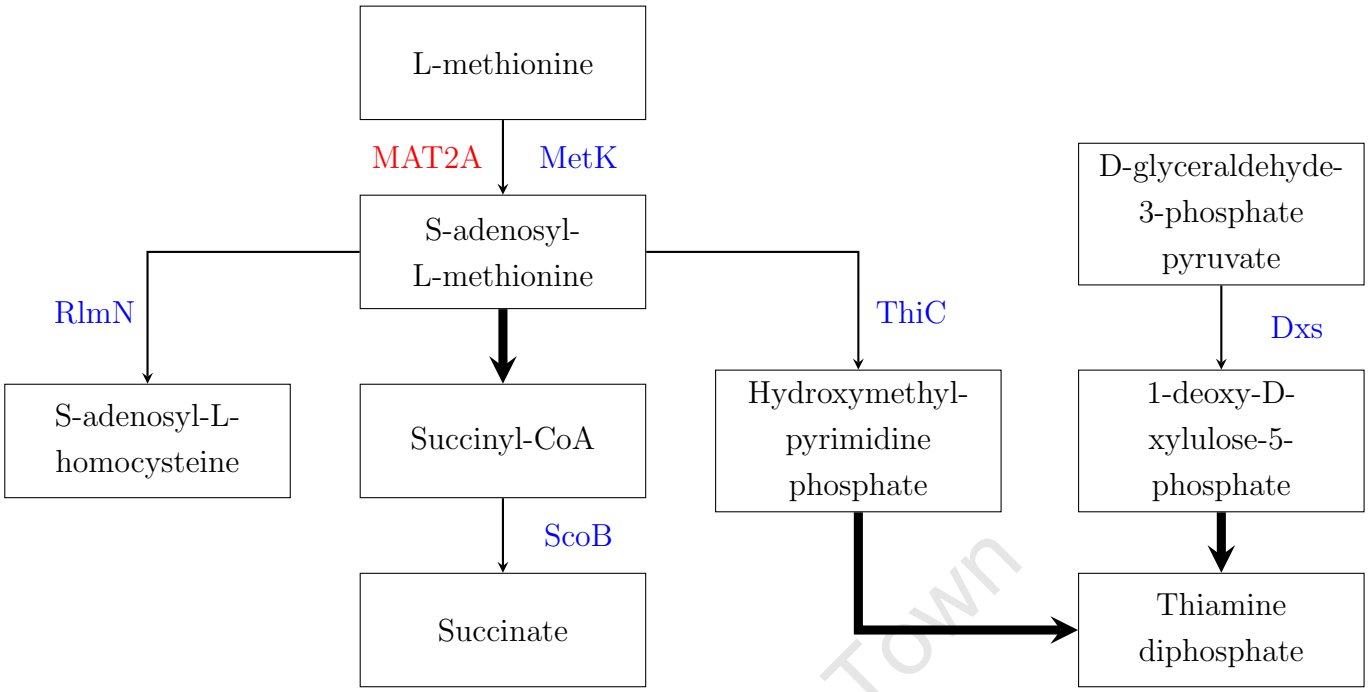


Figure 3.3: Interaction involving MAT2A. Thin arrow indicates direct interaction, whereas thick arrow indicates that there are several steps.

which is consumed during the biosynthesis of the thiazole moiety of thiamin. ScoB is the B subunit of probable succinyl-CoA:3-ketoacid-coenzyme A transferase.

We also looked at host-pathogen interactions where both proteins belong to the same pathway but *Mtb* does not have the equivalent of the human protein in its pathway. The human protein GOT2 (P00505), an aspartate aminotransferase, mitochondrial, interacts with the *Mtb* protein Rv0458 (P63937), which is a probable aldehyde hydrogenase. Both proteins are involved in arginine and proline metabolism but the *Mtb* pathway lacks the aspartate aminotransferase protein. Two interactions, the first between human protein SUCLA2 (Q9P2R7, EC:6.2.1.4) and *Mtb* SucD (P71558, EC:6.2.1.5) and the second between human protein SUCLG1 (P53597, EC:6.2.1.4, 6.2.1.5) and *Mtb* SucC (P71559, EC:6.2.1.5), have all proteins belonging to the citrate cycle (TCA) pathway. The two human proteins catalyse the ligation of succinate and CoA to form succinyl-CoA, the ligation is ATP-dependent for SUCLA2 and ATP- or GTP-dependent for SUCLG1. However, *Mtb* only possesses the enzyme with EC number 6.2.1.5, which is the succinate–CoA ligase (ADP-forming). These could represent examples of Mtb possibly using host proteins or products of their reactions to complete its pathways.

Interestingly, the first pathway retrieved while searching the human "database-array" proteins against KEGG is the tuberculosis pathway. 8 proteins, namely CTSD, HSPD1, JAK1,

NFKB1, RAB5A, RAB5C, CALM1 (P62158) and CD74, were mapped onto the tuberculosis pathway. In particular, CTSD, RAB5A, RAB5C and CALM1 are all involved in phagosome maturation which is blocked by *Mtb*. CALM1 is a Ca^{2+}-dependent effector protein necessary for the activation of CaMKII, which is required for the recruitment of EEA1 to the phagosome. RAB5A and RAB5C are GTPases necessary for phagosome maturation through the recruitment of a large number of effector proteins. Lastly, CTSD or cathepsin D is a hydrolytic protease secreted by the phagolysosome for bacterial degradation. Figure 3.4 represents the proteins highlighted on the tuberculosis pathway from KEGG.

While searching the *Mtb* proteins "interolog-array" or "database-array" against KEGG, one protein from "interolog-array", namely GroEL (P0A520), belongs to the tuberculosis pathway. GroEL (HSP65, highlighted in Figure 3.4) is a 60 kDa chaperonin 2 protein known to interact with TLR4. In the "interolog-array" data, it was predicted to interact with three human proteins: CLPP (Q16740), a putative ATP-dependent Clp protease proteolytic subunit, mitochondrial, TUFM (P49411), an elongation factor Tu, mitochondrial, and PRPS2 (P11908), which is a ribose-phosphate pyrophosphokinase 2. Even if the other *Mtb* proteins did not map to the tuberculosis pathway, some of them have biological processes that are present in the known *Mtb* proteins interacting with human proteins (Table 3.5). Metabolic pathways, biosynthesis of secondary metabolites and microbial metabolism in diverse environments are the top three pathways the "interolog-array" and "database-array" proteins map to.

3.6 Predicted interactions and drug targets

One of the main reason to study host-pathogen protein interactions is the possibility of finding targets to create new drugs. This is particulary important in the case of tuberculosis because of the existence of drug resistance.

A list of 881 potential drug target proteins was computationally predicted for the *Mtb* network in a separate study in the laboratory [84]. These are important proteins in the *Mtb* functional network since they are responsible for several indirect functional connections between other proteins in network. 878 out of 881 proteins were present in our *Mtb* network. We used this list to identify potential drug targets in the *Mtb* proteins predicted to interact with human proteins and we used the Fisher's exact test to determine whether the predicted list of proteins contains more drug targets than by chance. Table 3.8 presents the number of drug targets in each *Mtb* protein list with the p-value given by the Fisher's exact test. Taking a cut-off of 0.05 for the p-value, the "interolog-array" and "database-array" protein lists con-

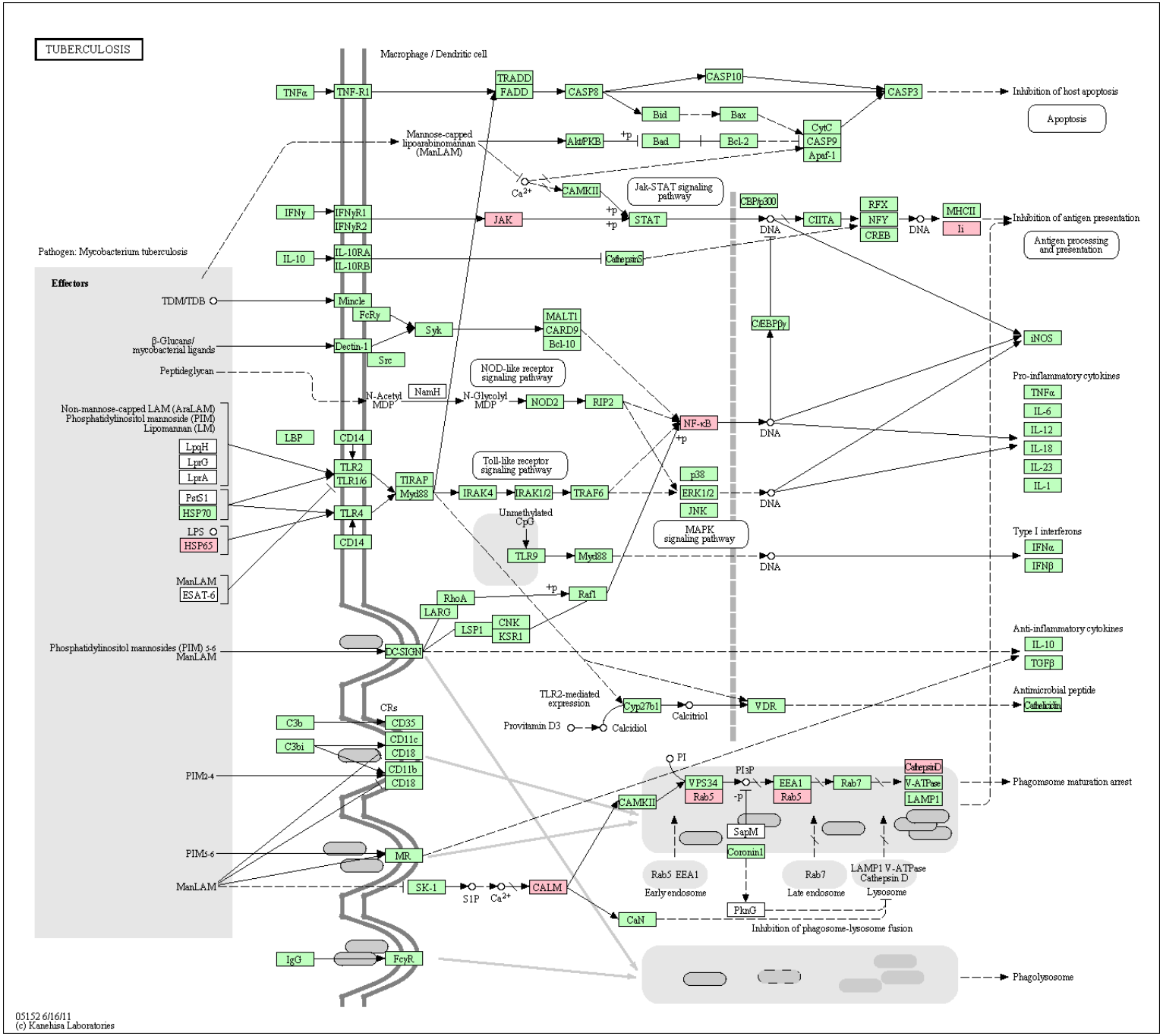


Figure 3.4: The "database-array" proteins overlayed on the tuberculosis pathway. Pink background represents the proteins present in the "database-array" list.

tain more drug targets than would be expected by chance, with p-values of 2.41×10^{-6} and 3.80×10^{-5}, respectively. Figures 3.5 and 3.6 highlight the *Mtb* proteins drug targets in the "interolog-array" and "database-array" interactions, respectively. 5 of the human proteins from the "database-array" interactions that mapped onto the tuberculosis pathway interact with *Mtb* proteins predicted to be drug target. In particular, the 4 *Mtb* proteins interacting with CD74 are all predicted drug targets.

Mtb protein list	Number of drug targets	p-value
Interolog-known	1/3	0.52
Interolog-array	25/47	2.41×10^{-6}
Database-array	25/53	3.80×10^{-5}

Table 3.8: Number of drug targets and p-value for each *Mtb* protein list.

3.7 Discussion and conclusion

In this chapter, we analysed the predicted host-pathogen interactions to see if they are likely to take place. In order to do so, we first looked at the subcellular localization of the proteins using the GO cellular component. We noted that human proteins known to interact with *Mtb* proteins are secreted or located in the membrane. Likewise, *Mtb* proteins known to interact with human proteins are secreted, located on the cell wall or cell membrane. Indeed, the known human-*Mtb* interactions come from studies focusing on the recognition of *Mtb* proteins by human receptors, thus the surface localization of the proteins. On the other hand, human proteins predicted to interact with *Mtb* proteins are mostly annotated as intracellular, though several human proteins from the "interolog-database" are annotated as membrane or secreted. *Mtb* proteins predicted to interact with human proteins are mostly found on the cell membrane, cell wall or are secreted, only a few are located in the cytoplasm. The subcellular locations of the predicted proteins reflect the fact that the bacteria is located within the host cells and suggests that all of these predictions are potentially feasible.

We also performed a GO biological process term enrichment on the predicted proteins using the DAVID web tools. The human "interolog-array" proteins predicted to interact with *Mtb* proteins were enriched in nitrogen compound biosynthetic process, oxoacid metabolic process, carboxylic acid metabolic process and nucleobase, nucleoside and nucleotide metabolic process. The human "database-array" proteins predicted to interact with *Mtb* were enriched in biological terms positive regulation of cellular process, regulation of apoptosis, regulation

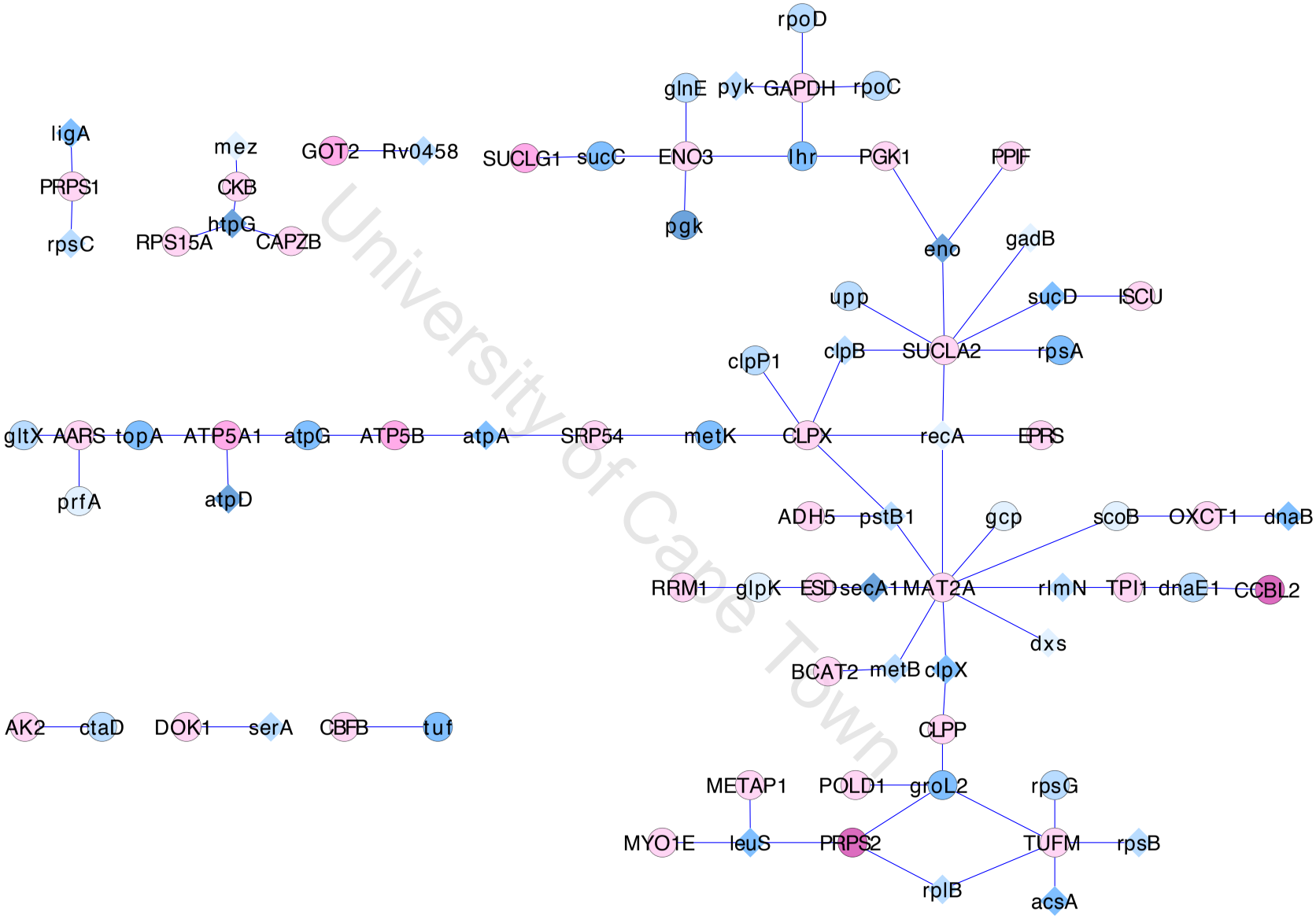


Figure 3.5: Drug targets in the "interolog-array" network. Drug targets have a diamond shape.

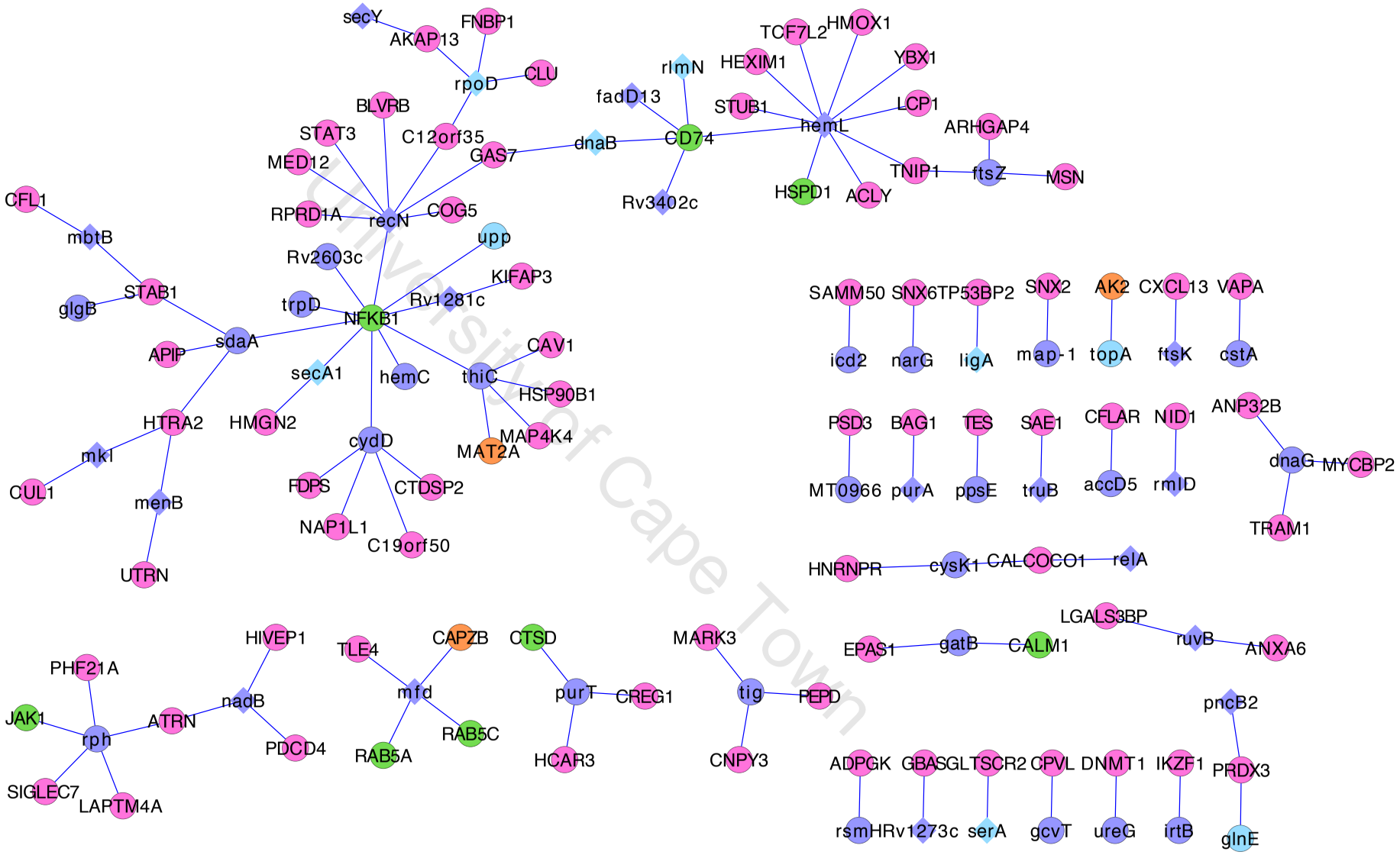


Figure 3.6: Drug targets in the "database-array" network. Drug targets have a diamond shape.

of programmed cell death and regulation of cell death. In particular, we had negative regulation of apoptosis, negative programmed cell death and negative regulation of cell death. All of the three biological processes have the same 11 proteins, listed in Table 3.9 and they interact with 21 bacterial proteins. Among these 11 human proteins, 5 interact with human anti-apoptosis genes that are known to be upregulated during *Mtb* infection, namely MCL-1 (Q07820), Bfl1 (Q16548), FLIP (Q8NFZ5) and BCL-2 (P10415) [7, 19]. *Mtb* might therefore interact with these proteins to inhibit apoptosis. These results are in accordance with the fact that *Mtb* inhibits apoptosis of the host macrophage at the early stage of infection to be able to survive and persist in the macrophages. Three *Mtb* proteins are known to inhibit apoptosis during early infection: NuoG (P95175), which encodes one subunit of the type I NADH dehydrogenase, PknE (P72001), a serine threonine kinase and SodA, a superoxide dismutase which is not present in our *Mtb* network. However, SodA is secreted by SecA2 (P66785), a bacterial secretion system [19]. We found that SecA2 interacts with SecA1, which in turn interacts with NFKB1, one of the human protein interacting with human anti-apoptosis gene.

Uniprot Acc	Protein name	Gene name
Q99933	BAG family molecular chaperone regulator 1 (Bcl-2-associated athanogene 1)	BAG1
O15519	CASP8 and FADD-like apoptosis regulator	CFLAR
P10809	60 kDa heat shock protein, mitochondrial	HSPD1
P10909	Clusterin	CLU
P23528	Cofilin-1 (18 kDa phosphoprotein)	CFL1
P14625	Endoplasmin (94 kDa glucose-regulated protein) (GRP-94)	HSP90B1
P04233	HLA class II histocompatibility antigen gamma chain	CD74
P09601	Heme oxygenase 1	HMOX1
P19838	Nuclear factor NF-kappa-B p105 subunit	NFKB1
P30048	Thioredoxin-dependent peroxide reductase, mitochondrial	PRDX3
Q9NQB0	Transcription factor 7-like 2 (T-cell-specific transcription factor 4)	TCF7L2

Table 3.9: Human proteins involved in apoptosis

When analysing the pathways the predicted proteins belong to, we found that the three *Mtb* proteins MetK, RlmN and ThiC interacting with the human MAT2A protein use L-methionine or S-adenosyl-L-methionine (SAM). Methionine is a key amino acid for protein structure and for metabolism [20] and is mostly used to synthesize SAM, which is the principal biological methyl donor [81]. *Mtb* was shown to import methionine from the host [88]. However, the methionine transporter is as yet unknown. Our result may suggest that *Mtb*

uses methionine imported from the host to produce S-adenosyl-L-methionine. Mapping of the human "database-array" proteins onto KEGG also retrieved the tuberculosis pathway. In particular, 5 proteins were involved in phagolysosome fusion, which is blocked by *Mtb* during infection.

Guérin and de Chastellier [58] proposed that *M. avium* impairs the movement of early endosomes by disrupting the actin filament network. This reduces the probability of fusion and intermingling of membrane and contents between the endosomes and phagosomes containing this pathogenic mycobacterium. They hypothesise that *M. avium* disrupts the actin filament network by direct interaction of bacterial constituents with proteins of, or associated with, the phagosome membrane being involved in depolymerization/polymerization of actin filaments. Furthermore, they mentioned several host proteins, namely the actin motor myosin-I, the actin binding proteins ezrin and moesin, the small GTPase of the Rho family RhoD, and the bacterial macrophage-induced gene (mig) protein as potential candidates for the disruption of the actin filaments. We predicted that both human proteins moesin (P26038) and ARHGAP4 (P98171), which is a Rho GTPase-activating protein 4, interact with the bacterial protein FtsZ (P64170), a cell division protein and prokaryotic homolog of tubulin. An interaction between human MYO1E (Q12965), an unconventional myosin-I which serves in intracellular movements, and LeuS (P67510), a leucyl-tRNA synthetase, was also predicted.

The bacterial protein HtpG is a heat shock protein belonging to the 90-kDa molecular chaperone family. Heat shock proteins (hsp) are major antigens in a number of infectious diseases. In particular, immune response to members of the Hsp90 family has been observed in schistosomiasis, malaria, Chaga's disease and candidiasis. HtpG was predicted to interact with the human protein CAPZB, which is a a F-actin-capping protein subunit beta. F-actin-capping proteins bind in a Ca^{2+}-independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends. CAPZB has actin filament capping, which is a child of negative regulation of actin filament depolymerization, as GO biological process. CAPZB was also predicted to interact with the bacterial protein mfd (P64326), which is a transcription-repair-coupling factor.

Macrophages produce reactive nitrogen and oxygen intermediate to kill *Mtb*. Human ADH5 (P11766) or GSNOR is a S-nitrosoglutathione reductase which regulates S-nitrosothiols (SNOs) and nitric oxide (NO) *in vivo* through catabolism of S-nitrosoglutathione [126], more precisely it blocks NO function by reducing S-nitrosoglutathione (GSNO) to NH3 [65]. It is therefore an important protein in the control of NO metabolism. ADH5 was predicted to interact with the *Mtb* pstB1 (O53832), which is a phosphate-import ATP-binding protein. The human protein EstD (P10768) is a S-formylglutathione hydrolase. It hydrolyses

S-formylglutathione to formate and glutathione, which is toxic to mycobacteria [38]. EstD was predicted to interact with two *Mtb* proteins: GlpK (O69664), a glycerol kinase and SecA1 (P0A5Y8), which is a protein translocase subunit SecA 1.

Interestingly, the "interolog-array" and "database-array" *Mtb* proteins contain more potential drug targets than due to chance alone. These proteins would therefore be of great importance for drug design provided that their role in host interactions were experimentally confirmed.

Inter-species protein interaction prediction using interologs is able to capture the fact that *Mtb* is inside the host during infection. However, the GO biological processes and the pathways are quite different depending on the initial data used to predict the interologs. Nonetheless, the results are complementary since they give insight into how *Mtb* might acquire nutrients and how it modulates the host response to its advantage.

Chapter 4

Conclusion

Host-pathogen protein interactions are essential for the pathogen to establish infection and to be able to survive inside the host. However, experimental studies of these interactions are scarce and computational methods are used instead to predict the interactions.

This project was set out in order to predict protein interactions between human and *Mtb*. We used computational methods based on interologs to predict them and overlaid the predicted interactions onto constructed human and *Mtb* protein interaction networks. The human protein interaction network was constructed by integrating data from Bossi and Lehner, REACTOME and STRING. On the other hand, the *Mtb* network combined data from STRING, gene expression and sequence data. We predicted interologs by using experimentally verified intra-species and inter-species interactions, that were furthermore filtered using expression data. We also collected already known host-pathogen interactions by literature mining and found 47 interactions, which were subsequently used as a filter for the interologs predicted using intra-species interactions to find the "interolog-known" interactions. We assessed the likelihood that the predicted interactions would occur *in vivo* by looking at their cellular components, biological processes and pathways. The *Mtb* proteins were located on the cell surface, whereas human proteins were mostly tagged as intracellular, showing the intracellular location of the bacteria. Analysis of the biological processes of the proteins suggests that the human "interolog-array" proteins are involved in facilitating the production of nitric oxide whereas the "database-array" proteins are related to negative regulation of apoptosis. Predicted host-pathogen interactions had very low function similarity score suggesting that interacting human and *Mtb* proteins perform different functions. Mapping the predicted interactions onto KEGG pathways revealed that some of the proteins are known to play a role in tuberculosis and that *Mtb* might hijack the host to acquire nutrients. We also found that the predicted *Mtb* proteins are enriched in predicted drug targets.

Human-*Mtb* interaction prediction is of great importance for understanding tuberculosis. Our work predicts interactions that could help understand the interplay between the host and pathogen and may prove useful for designing new drugs. However, all predictions should ideally be verified in the laboratory, and technologies are now available for experimental identification of inter-species protein interactions [111, 136]. In future work, in addition to experimental validation, the interologs method used here could be combined with other computational protein-protein interaction prediction methods to improve the results. These include the use of knowledge on interacting domains. Future work could also include applying the interologs method to predict host-pathogen interactions in other organisms.

References

[1] H. Adachi and M. Tsujimoto. FEEL-1, a novel scavenger receptor with *in vitro* bacteria-binding and angiogenesis-modulating activities. *The Journal of Biological Chemistry*, 277(37):34264–34270, September 2002.

[2] S. Ahmad. Pathogenesis, immunology, and diagnosis of latent *Mycobacterium tuberculosis* infection. *Clinical and Developmental Immunology*, 2011, 2010.

[3] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[4] E. Alix, S. Mukherjee, and C. R. Roy. Subversion of membrane transport pathways by vacuolar pathogens. *Journal of Cell Biology*, pages 61–70, 2012.

[5] F. I. Andersson, A. Tryggvesson, M. Sharon, A. V. Diemand, M. Classen, C. Best, R. Schmidt, J. Schelin, T. M. Stanne, B. Bukau, C. V. Robinson, S. Witt, A. Mogk, and A. K. Clarke. Structure and function of a novel type of ATP-dependent clp protease. *The Journal of Biological Chemistry*, 284(20):13519–13532, 2009.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[7] H. Ashida, H. Mimuro, M. Ogawa, T. Kobayashi, T. Sanada, M. Kim, and C. Sasakawa. Cell death and infection: A double-edged sword for host and pathogen survival. *Journal of Cell Biology*, 195(3), 2011.

[8] H. Bach, K. G. Papavinasasundaram, D. Wong, Z. Hmama, and Y. Av-Gay. *Mycobacterium tuberculosis* virulence is mediated by PtpA dephosphorylation of human vacuolar protein sorting 33B. *Cell Host Microbe*, 3:316–322, 2008.

[9] G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, and C. W. V. Hogue. BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Research*, 29(1):242–245, 2000.

[10] A. Baena and S. A. Porcelli. Evasion and subversion of antigen presentation by *Mycobacterium tuberculosis*. *Tissue Antigens*, 74(3):189–204, 2009.

[11] K. Bansal, S. R. Elluru, Y. Narayana, R. Chaturvedi, S. A. Patil, S. V. Kaveri, J. Bayry, and K. N. Balaji. PE_PGRS antigens of *Mycobacterium tuberculosis* induce maturation and activation of human dendritic cells. *The Journal of Immunology*, 184(7):3495–3504, 2010.

[12] K. Bansal, A. Y. Sinha, D. S. Ghorpade, S. K. Togarsimalemath, S. A. Patil, S. V. Kaveri, K. N. Balaji, and J. Bayry. Src homology 3-interacting domain of Rv1917c of *Mycobacterium tuberculosis* induces selective maturation of human dendritic cells by regulating PI3K-MAPK-NF-κ B signaling and drives Th2 immune responses. *Journal of Biological Chemistry*, 285(47):36511–36522, 2010.

[13] S. Banu, N. Honoré, B. Saint-Joanis, D. Philpott, M.-C. Prévost, and S. T. Cole. Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Molecular Microbiology*, 44(1):9–19, 2002.

[14] S. Basu, S. K. Pathak, A. Banerjee, S. Pathak, A. Bhattacharyya, Z. Yang, S. Talarico, M. Kundu, and J. Basu. Execution of macrophage apoptosis by PE_PGRS33 of *Mycobacterium tuberculosis* is mediated by Toll-like receptor 2-dependent release of tumor necrosis factor-α. *The Journal of Biological Chemistry*, 282:1039–1050, 2007.

[15] T. Berggard, S. Linse, and P. James. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7:2833–2842, 2007.

[16] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.

[17] C. Blaschke, R. Hoffmann, J. C. Oliveros, and A. Valencia. Extracting information automatically form biological literature. *Comparative and Functional Genomics*, 2(5):310–313, 2001.

[18] A. Bossi and B. Lehner. Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5:260, 2009.

[19] V. Briken and J. L. Miller. Living on the edge: inhibition of host cell apoptosis by *Mycobacterium tuberculosis*. *Future Microbiology*, 3:415–422, August 2008.

[20] J. T. Brosnan, M. E. Brosnan, R. F. Bertolo, and J. A. Brunton. Methionine: A metabolically unique amino acid. *Livestock Science*, 112:2–7, 2007.

[21] K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, May 2005.

[22] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

[23] F. Browne, H. Zheng, H. Wang, and F. Azuaje. From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. *Advances in Artificial Intelligence*, 2010, 2010.

[24] Y. Bulut, K. S. Michelsen, L. Hayrapetian, Y. Naiki, R. Spallek, M. Singh, and M. Arditi. *Mycobacterium tuberculosis* heat shock proteins use diverse toll-like receptor pathways to activate pro-inflammatory signals. *The Journal of Biological Chemistry*, 280(22):20961–20967, 2005.

[25] E. Camon, D. Barrell, C. Brooksbank, M. Magrane, and R. Apweiler. The Gene Ontology Annotation (GOA) project — application of GO in SWISS-PROT, TrEMBL and InterPro. *Comparative and Functional Genomics*, 4:71–74, 2003.

[26] J.-C. Camus, M. J. Pryor, C. Médigue, and S. T. Cole. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, 148(10):2967–2973, October 2002.

[27] R. A. Capaldi, B. Schulenberg, J. Murray, and R. Aggeler. Cross-linking and electron microscopy studies of the structure and functioning of the *Escherichia Coli* atp synthase. *The Journal of Experimental Biology*, 203:29–33, 2000.

[28] G. Cappelli, E. Volpe, M. Grassi, B. Liseo, V. Colizzi, and F. Mariani. Profiling of *Mycobacterium tuberculosis* gene expression during human macrophage infection: upregulation of the alternative sigma factor G, a group of transcriptional regulators, and proteins with unknown function. *Research in Microbiology*, 157(5):445–455, 2006.

[29] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, and S. Lewis. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.

[30] M. V. Carroll, R. B. Sim, F. Bigi, A. Jäkel, R. Antrobus, and D. A. Mitchell. Identification of four novel DC-SIGN ligands on *Mycobacterium bovis* BCG. *Protein Cell*, 1(9):859–870, 2010.

[31] D. Chaussabel, R. Semnani, M. McDowell, D. Sacks, A. Sher, and T. Nutman. Unique gene expression profiles of human macrophages and dendritic cells to phylogenitically distinct parasites. *Blood*, 102(2):672–681, 2003.

[32] L. Chen, X. He, X. Zhao, and J. Su. Interaction of *Mycobacterium tuberculosis* MPB64 protein with heat shock protein 40. *African Journal of Microbiology Research*, 5:394–398, 2011.

[33] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. B. III, F. Tekaia, K. Badcock, d. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, June 1998.

[34] H. L. Collins. The role of iron in infections with intracellular bacteria. *Immunology Letters*, 85(2):193–195, January 2003.

[35] D. V. Cousins, R. Bastidia, A. Cataldi, V. Quse, S. Redrobe, S. Dow, P. Duignan, A. Murray, C. Dupont, N. Ahmed, D. M. Collins, W. R. Butler, D. Dawson, D. Rodriguez, J. Loureiro, M. I. Romano, A. Alito, M. Zumarraga, and A. Bernardelli. Tuberculosis in seals caused by a novel member of the Mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 53(5):1305–1314, 2003.

[36] T. Cui, L. Zhang, X. Wang, and Z.-G. He. Uncovering new signaling proteins and potential drug targets through the interactome analysis of mycobacterium tuberculosis. *BMC Genomics*, 10:118, 2009.

[37] F. P. Davis, D. T. Barkan, N. Eswar, J. H. McKerrow, and A. Sali. Host-pathogen protein interactions predicted by comparative modeling. *Protein Science*, 16(12):2585–2596, 2007.

[38] Y. K. Dayaram, M. T. Talaue, N. D. Connell, and V. Venketaraman. Characterization of a glutathione metabolic mutant of *Mycobacterium tuberculosis* and its resistance to glutathione and nitrosoglutathione. *Journal of Bacteriology*, 188(4):1364–1372, 2006.

[39] J. M. Doolittle and S. M. Gomez. Structural similarity-based predictions of protein interactions between HIV-1 and *Homo sapiens*. *Virology Journal*, 7:82, 2010.

[40] M. G. Drage, N. D. Pecora, A. G. Hise, M. Febbraio, R. L. Silverstein, D. T. Golenbock, W. H. Boom, and C. V. Harding. Tlr2 and its co-receptors determine responses of macrophages and dendritic cells to lipoproteins of *Mycobacterium tuberculosis*. *Cellular Immunology*, 258(1):29–37, Apr 2009.

[41] T. Driscoll, M. D. Dyer, T. M. Murali, and B. W. Sobral. PIG—the pathogen interaction gateway. *Nucleic Acids Research*, 37(suppl 1):D647–D650, 2009.

[42] R. G. Ducati, A. Ruffino-Netto, L. A. Basso, and D. S. Santos. The resumption of consumption – a review on tuberculosis. *Memórias do Instituto Oswaldo Cruz*, 101(7):697–714, 2006.

[43] M. D. Dyer, T. M. Murali, and B. W. Sobral. Computational prediction of host-pathogen protein protein interactions. *Bioinformatics*, 23(13):i159–166, 2007.

[44] M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali, and B. W. Sobral. The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *Plos One*, 5(8):e12089, 2010.

[45] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.

[46] P. Evans, W. Dampier, L. Ungar, and A. Tozeren. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Medical Genomics*, 2:27, 2009.

[47] G. Ferrari, H. Langen, M. Naito, and J. Pieters. A coat protein on phagosomes involved in the intracellular survival of mycobacteria. *Cell*, 97(4):435–447, May 1999.

[48] R. D. Fleischmann, D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. J. Jr, J. C. Venter, and C. M. Fraser. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *Journal of Bacteriology*, 184(19):5479–5490, 2002.

[49] C. V. Forst. Host-pathogen systems biology. *Drug Discovery Today*, 11(5–6):220–227, March 2006.

[50] C. C. Frota, D. M. Hunt, R. S. Buxton, L. Rickman, J. Hinds, K. Kremer, D. van Soolingen, and M. J. Colston. Genome structure in the vole bacillus, *Mycobacterium microti*, a member of the *Mycobacterium tuberculosis* complex with a low virulence for humans. *Microbiology*, 150(5):1519–1527, 2004.

[51] S. Gagneux and P. M. Small. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *The Lancet Infectious Diseases*, 7(5):328–337, 2007.

[52] A. J. Gehring, K. M. Dobos, J. T. Belisle, C. V. Harding, and W. H. Boom. *Mycobacterium tuberculosis* LprG (Rv1411c): A novel TLR-2 ligand that inhibits human macrophage class II MHC antigen processing. *Journal of Immunology*, 173(4):2660–2668, 2004.

[53] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolma, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. F. Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, 2003.

[54] S. E. Girardin, I. G. Boneca, J. Viala, M. Chamaillard, A. Labigne, G. Thomas, D. J. Philpott, and P. J. Sansonetti. Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. *The Journal of Biological Chemistry*, 278(11):8869–8872, Mar 2003.

[55] K.-I. Goh, B. Kahng, and D. Kim. Graph theoritic analysis of protein interaction networks of eukaryotes. *Physica A: Statistical Mechanics and its Applications*, 357(3–4):501–512, 2005.

[56] A. Gordon, P. Hart, and M. Young. Ammonia inhibits phagosom-lysosome fusion in macrophages. *Nature*, 286:79–80, 1980.

[57] M. Goren, P. D. Hart, M. Young, and J. Armstrong. Prevention of phagosome-lysosome fusion in cultured macrophages by sulfatides of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, 73(7):2510–2514, 1976.

[58] I. Guérin and C. de Chastellier. Disruption of the actin filament network affects delivery of endocytic contents marker to phagosomes with early endosome characteristics: The case of phagosomes with pathogenic mycobacteria. *European Journal of Cell Biology*, 79:735–749, 2000.

[59] C. V. Harding and W. H. Boom. Regulation of antigen presentation by *Mycobacterium tuberculosis*: a role for Toll-like receptors. *Nature Reviews Microbiology*, 8(4):296–307, Apr 2010.

[60] G. Hetland and H. Wiker. Antigen 85C on *Mycobacterium bovis*, BCG and *M. Tuberculosis* promotes monocyte-CR3-mediated uptake of microbeads coated with mycobacterial products. *Immunology*, 82(3):445–449, 1994.

[61] E. W. Hewitt. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, 110(2):163–169, October 2003.

[62] R. Himmelreich, H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research*, 24(22):4420–4449, 1996.

[63] D. A. Hosack, J. Glynn Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10):R70, 2003.

[64] A. Jain and P. Dixit. Multidrug-resistant to extensively drug resistant tuberculosis: what is next? *Journal of Biosciences.*, 33(4):605–616, November 2008.

[65] D. Jensen, G. Belka, and G. D. Bois. S-nitrosoglutathione is a substrate for rat alcohol dehydrogenase class III isoenzyme. *Biochemical Journal*, 331(2):659–668, 1998.

[66] S. Józefowski, A. Sobota, and K. Kwiatkowska. How *Mycobacterium tuberculosis* subverts host immune responses. *BioEssays*, 30(10):943–954, October 2008.

[67] S.-B. Jung, C.-S. Yang, J.-S. Lee, A.-R. Shin, S.-S. Jung, J. W. Son, C. V. Harding, H.-J. Kim, J.-K. Park, T.-H. Paik, C.-H. Song, and E.-K. Jo. The mycobacterial 38-kilodalton glycolipoprotein antigen activates the mitogen- activated protein kinase pathway and release of proinflammatory cytokines through toll-like receptors 2 and 4 in human monocytes. *Infection and Immunity*, 74(5):2686–2696, 2006.

[68] A. Karboul, A. Mazza, N. C. G. van Pittius, J. L. Ho, R. Brousseau, and H. Mardassi. Frequent homologous recombination events in *Mycobacterium tuberculosis* PE/PPE multigene families: potential role in antigenic variability. *Journal of Bacteriology*, 190(23):7838–7846, 2008.

[69] N. Khan, K. Alam, S. C. Mande, V. L. Valluri, S. E. Hasnain, and S. Mukhopadhyay. *Mycobacterium tuberculosis* heat shock protein 60 modulates immune response to PPD by manipulating the surface expression of TLR2 on macrophages. *Cellular Microbiology*, 10(8):1711–1722, 2008.

[70] J. Kleinnijenhuis, M. Oosting, L. A. B. Joosten, M. G. Netea, and R. V. Crevel. Innate immune recognition of *Mycobacterium tuberculosis*. *Clinical and Developmental Immunology*, 2011.

[71] W. Kress, Ẑeljka Maglica, and E. Weber-Ban. Clp chaperone-protease: structure and function. *Research in Microbiology*, 160:618–628, 2009.

[72] O. Krishnadev, S. Bisht, and N. Srinivasan. Prediction of protein-protein interactions between human host and two mycobacterial organisms. *International Journal of Knowledge Discovery in Bioinformatics*, 1(1):1–13, 2010.

[73] G. P. Kubica, T. H. Kim, and F. P. Dunbar. Designation of strain H37Rv as the Neotype of *Mycobacterium tuberculosis*. *International Journal of Systematic Bacteriology*, 22(2):99–106, April 1972.

[74] K. Kumar, M. Tharad, S. Ganapathy, G. Ram, A. Narayan, J. A. Khan, R. Pratap, A. Ghosh, S. K. Samuchiwal, S. Kumar, K. Bhalla, D. Gupta, K. Natarajan, Y. Singh, and A. Ranganathan. Phenylalanine-rich peptides potently bind ESAT6, a virulence determinant of *Mycobacterium tuberculosis*, and concurrently affect the pathogen's growth. *PLoS ONE*, 4(11):e7615, 2009.

[75] R. Kumar and B. Nanduri. HPIDB - a unified resource for host-pathogen interactions. *BMC Bioinformatics*, 11(Suppl 6):S16, 2010.

[76] J. Lee, M. Hartman, and H. Kornfeld. Macrophage apoptosis in tuberculosis. *Yonsei Medical Journal*, 50(1):1–11, February 2009.

[77] S.-A. Lee, C. hsiung Chan, C.-H. Tsai, J.-M. Lai, F.-S. Wang, C.-Y. Kao, and C.-Y. F. Huang. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*, 9:S11, 2008.

[78] B. Lehner and A. G. Fraser. A first-draft human protein-interaction map. *Genome Biology*, 5(9):R63, Aug 2004.

[79] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J.-F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Z. G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. S. A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–543, January 2004.

[80] C. Loeuillet, F. Martinon, C. Perez, M. Munoz, M. Thome, and P. R. Meylan. *Mycobacterium tuberculosis* subverts innate immunity to evade specific effectors. *The Journal of Immunology*, 177:6245–6255, 2006.

[81] S. C. Lu. S-Adenosylmethionine. *The International Journal of Biochemistry and Cell Biology*, 32:391–395, 2000.

[82] M. Luo, E. A. Fadeev, and J. T. Groves. Mycobactin-mediated iron acquisition within macrophages. *Nature Chemical Biology*, 1:149–153, 2005.

[83] Z. A. Malik, G. M. Denning, and D. J. Kusner. Inhibition of Ca^{2+} signaling by *Mycobacterium tuberculosis* associated with reduced phagosome-lysosome fusion and increased survival within human macrophages. *The Journal of Experimental Medicine*, 191(2):287–302, 2000.

[84] G. K. Mazandu and N. J. Mulder. Generation and analysis of large-scale data-driven *Mycobacterium tuberculosis* functional networks for drug target identification. *Advances in Bioinformatics*, 2011:Article ID 801478, 2011.

[85] G. K. Mazandu and N. J. Mulder. Scoring protein relationships in functional interaction networks predicted from sequence data. *PLoS ONE*, 6(4):e18607, 2011.

[86] G. K. Mazandu and N. J. Mulder. A topology-based metric for measuring term similarity in the Gene Ontology. *Advances in Bioinformatics*, 2012:Ariticle ID 975783, 2012.

[87] G. K. Mazandu, K. Opap, and N. J. Mulder. Contribution of microarray data to the advancement of knowledge of the *Mycobacterium tuberculosis* interactome: Use of the random partial least squares approach. *Infection, Genetics and Evolution*, 11(4):725–733, 2011.

[88] R. A. McAdam, T. R. Weisbord, J. Martin, J. D. Scuderi, A. M. Brown, J. D. Cirillo, B. R. Bloom, and W. R. J. Jr. In vivo growth characteristics of leucine and methionine auxotrophic mutants of mycobacterium bovis BCG generated by transposon mutagenesis. *Infection and Immunity*, 63(3):1004–1012, 1995.

[89] G. Migliori, G. D. Iaco, G. Besozzi, R. Centis, and D. Cirillo. First tuberculosis cases in Italy resistant to all tested drugs. *Eurosurveillance*, 12(20), 2007.

[90] B. B. Mishra, P. Moura-Alves, A. Sonawane, N. Hacohen, G. Griffiths, L. F. Moita, and E. Anes. *Mycobacterium tuberculosis* protein ESAT-6 is a potent activator of the NLRP3/ASC inflammasome. *Cellular Microbiology*, 12(8):1046–1063, 2010.

[91] S. L. Mueller-Ortiz, A. R. Wanger, and S. J. Norris. Mycobacterial protein HbhA binds human complement component C3. *Infection and Immunity*, 69(12):7501–7511, 2001.

[92] S. Nair, A. D. Pandey, and S. Mukhopadhyay. The PPE18 protein of *Mycobacterium tuberculosis* inhibits NF-κB/rel-mediated proinflammatory cytokine production by upregulating and phosphorylating suppressor of cytokine signaling 3 protein. *The Journal of Immunology*, 186(9):5413–5424, 2011.

[93] V. Navratil, B. de Chassey, L. Meyniel, S. Delmotte, C. Gautier, P. André, V. Lotteau, and C. Rabourdin-Combe. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus–host interaction networks. *Nucleic Acids Research*, 37:D661–D668, 2009.

[94] A. G. Nerlich, C. J. Haas, A. Zink, U. Szeimes, and H. G. Hagedorn. Molecular evidence for tuberculosis in an ancient Egyptian mummy. *The Lancet*, 350(9088):1404, 1997.

[95] W. H. Organization. *Treatment of tuberculosis: guidelines – 4th ed.* WHO Press, 2010.

[96] W. H. Organization. Global tuberculosis control: WHO report 2011., 2011.

[97] T. Parish. Starvation survival response of *Mycobacterium tuberculosis*. *Journal of Bacteriology*, 185(22):6702–6706, November 2003.

[98] P. Peake, A. Gooley, and W. J. Britton. Mechanism of interaction of the 85B secreted protein of *Mycobacterium bovis* with fibronectin. *Infection and Immunity*, 61(11):4828–4834, 1993.

[99] N. D. Pecora, A. J. Gehring, D. H. Canaday, W. H. Boom, and C. V. Harding. *Mycobacterium tuberculosis* LprA is a lipoprotein agonist of TLR2 that regulates innate immunity and APC function. *Journal of Immunology*, 177(1):422–429, 2006.

[100] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[101] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. B. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. N. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, Oct 2003.

[102] M. Persico, A. Ceol, C. Gavrila, R. Hoffmann, A. Florio, and G. Cesareni. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 6 Suppl 4:S21, Dec 2005.

[103] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcao, and F. M. Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, 2008.

[104] S. Pitarque, J.-L. Herrmann, J.-L. Duteyrat, M. Jackson, G. R. Stewart, F. Lecointe, B. Payre, O. Schwartz, D. B. Young, G. Marchal, P. H. Lagrange, G. Puzo, B. Gicquel, J. Nigou, and O. Neyrolles. Deciphering the molecular bases of *Mycobacterium tuberculosis* binding to the lectin DC-SIGN reveals an underestimated complexity. *Biochemical Journal*, 392(3):615–624, 2005.

[105] T. P. Primm, S. J. Andersen, V. Mizrahi, D. Avarbock, H. Rubin, and I. Clifton E. Barry. The stringent response of *Mycobacterium tuberculosis* is required for long-term survival. *Journal of Bacteriology*, 182(17):4889–4898, 2000.

[106] H. Rachman, M. Strong, T. Ulrichs, L. Grode, J. Schuchhardt, H. Mollenkopf, G. Kosmiadi, D. Eisenberg, and S. Kaufmann. Unique transcriptome signature of mycobacterium tuberculosis in pulmonary tuberculosis. *Infection and Immunity*, 74(2):1233–1242, 2006.

[107] A. Ragas, L. Roussel, G. Puzo, and M. Rivière. The mycobacterium tuberculosis cell-surface glycoprotein apa as a potential adhesin to colonize target cells via the innate immune system pulmonary c-type lectin surfactant protein a. *The Journal of Biological Chemistry*, 282(8):5133–5142, Feb 2007.

[108] S. Ragno, M. Romano, S. Howell, D. Pappin, P. Jenner, and M. Colston. Changes in gene expression in macrophages infected with *Mycobacterium tuberculosis*: a combined transcriptomic and proteomic approach. *Immunology*, 104(1):99–108, 2001.

[109] A. K. Ramani, R. C. Bunescu, R. J. Mooney, and E. M. Marcotte. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):R40, Apr 2005.

[110] C. Ratledge. Iron, mycobacteria and tuberculosis. *Tuberculosis*, 84(1–2):110–130, 2004.

[111] K. D. Rodland, J. N. Adkins, C. Ansong, S. Chowdhury, N. P. Manes, L. Shi, H. Yoon, R. D. Smith, and F. Heffron. Use of high-throughput mass spectrometry to elucidate host-pathogen interactions in *Salmonella*. *Future Microbiology*, 3(6):625–634, 2008.

[112] C. Rosales. *Molecular mechanisms of phagocytosis*. Eurekah.Com Inc, 2005.

[113] K. Rowland and R. Guthmann. How should we manage a patient with a positive PPD and prior BCG vaccination? *The Journal of family practice*, 55(8):718–720, 2006.

[114] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005.

[115] C. M. Sassetti, D. H. Boyd, and E. J. Rubin. Comprehensive identification of conditionally essential genes in mycobacteria. *Proceedings of the National Academy of Sciences*, 98(22):12712–12717, 2001.

[116] C. M. Sassetti, D. H. Boyd, and E. J. Rubin. Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, 48(1):77–84, 2003.

[117] U. E. Schaible and S. H. Kaufmann. A nutritive view on the host-pathogen interplay. *TRENDS in Microbiology*, 13(8):373–380, August 2005.

[118] L. S. Schlesinger. Macrophage phagocytosis of virulent but not attenatuated strains of *Mycobacterium tuberculosis* is mediated by mannose receptors in addition to complement receptors. *The Journal of Immunology*, 150(7):2920–2930, 1993.

[119] N. W. Schluger and W. N. Rom. The host immune response to tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, 157:679–691, 1998.

[120] G. Schumann, S. Schleier, I. Rosenkrands, N. Nehmann, S. Hälbich, P. F. Zipfel, M. I. de Jonge, S. T. Cole, T. Munder, and U. Möllmann. *Mycobacterium tuberculosis* secreted protein ESAT-6 interacts with the human protein syntenin-1. *Central european journal of biology*, 1(2):183–202, 2006.

[121] S. K. Sharma and A. Mohan. Extrapulmonary tuberculosis. *The Indian Journal of Medical Research*, 120(4):316–353, 2004.

[122] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, 3(4):e43, 2007.

[123] C. D. Sohaskey and L. G. Wayne. Role of *narK2X* and *narGHJI* in hypoxic upregulation of nitrate reduction by *Mycobacterium tuberculosis*. *Journal of Bacteriology*, 185(24):7247–7256, 2003.

[124] C.-H. Song, J.-S. Lee, S.-H. Lee, K. Lim, H.-J. Kim, J.-K. Park, T.-H. Paik, and E.-K. Jo. Tumor necrosis factor-α, interleukin-10, and monocyte chemotactic protein-1 by *Mycobacterium tuberculosis* h37Rv-infected human monocytes. *Journal of Clinical Immunology*, 23(3):194–201, 2003.

[125] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde,

E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, Sep 2005.

[126] X. Sun, J. W. F. Wasley, J. Qiu, J. P. Blonder, A. M. Stout, L. S. Green, S. A. Strong, D. B. Colagiovanni, J. P. Richards, S. C. Mutka, L. Chuna, and G. J. Rosenthal. Discovery of S-Nitrosoglutathione reductase inhibitors: potential agents for the treatment of asthma and other inflammatory disease. *ACS Medicinal Chemistry Letters*, 2(5):402–406, 2011.

[127] B. Suter, S. Kittanakom, and I. Stagljar. Two-hybrid technologies in proteomics research. *Current Opinion in Biotechnology*, 19(4):316–323, 2008.

[128] L. Tailleux, S. Waddell, M. Pelizzola, A. Mortellaro, M. Withers, A. Tanne, P. Castagnoli, B. Gicquel, N. Stoker, P. Butcher, M. Foti, and O. Neyrolles. Probing host pathogen cross-talk by transcriptional profiling of both *Mycobacterium tuberculosis* and infected human dendritic cells and macrophages. *Plos One*, 3(1):e1403, 2008.

[129] K. Takahashi, M. Inuzuka, and T. Ingi. Cellular signaling mediated by calphoglin-induced activation of IPP and PGM. *Biochemical and Biophysical Research Communications*, 325(1):203–214, December 2004.

[130] O. Takeuchi and S. Akira. Pattern recognition receptors and inflammation. *Cell*, 140:805–820, 2010.

[131] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman. Prediction of interactions between HIV-1 and human proteins by information integration. *Pacific Symposium on Biocomputing*, 14:516–527, 2009.

[132] Z. F. Udwadia, R. A. Amale, K. K. Ajbani, and C. Rodrigues. Totally drug-resistant tuberculosis in india. *Clinical Infectious Diseases*, 54(4):579–581, 2011.

[133] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisae*. *Nature*, 403:623–627, 2000.

[134] S. Valway, M. Sanchez, T. Shinnick, I. Orme, T. Agerton, D. Hoy, J. Jones, H. Westmoreland, and I. Onorato. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. *The New England Journal of Medicine*, 338(10):633–639, March 1998.

[135] N. van der Wel, D. Hava, D. Houben, D. Fluitsma, M. van Zon, J. Pierson, M. Brenner, and P. J. Peters. *Mycobacterium tuberculosis* and M. leprae translocate from the phagolysosome to the cytosol in myeloid cells. *Cell*, 129(7):1287–1298, 2007.

[136] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995.

[137] A. Veleyati, M. Masjedi, P. Farnia, P. Tabarsi, J. Ghanavi, A. Ziazarifi, and S. Hoffner. Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in iran. *Chest*, 136(2):420–425, 2009.

[138] I. Vergne, J. Chua, and V. Deretic. Tuberculosis toxin blocking phagosome maturation inhibits a novel Ca^{2+}/Calmodulin-PI3K hVPS34 cascade. *Journal of Experimental Medicine*, 198(4):653–659, August 2003.

[139] I. Vergne, J. Chua, H.-H. Lee, M. Lucas, J. Belisle, and V. Deretic. Mechanism of phagolysosome biogenesis block by viable *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, 102(11):4033–4038, 2005.

[140] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.

[141] A. J. M. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287(5450):116–122, 2000.

[142] Y. Wang, T. Cui, C. Zhang, M. Yang, Y. Huang, W. Li, L. Zhang, C. Gao, Y. He, Y. Li, F. Huang, J. Zeng, C. Huang, Q. Yang, Y. Tian, C. Zhao, H. Chen, H. Zhang, and Z. He. Global protein-protein interaction network in the human pathogen *Mycobacterium tuberculosis* H37Rv. *Journal of Proteome Research*, 9(12):6665–77, 2010.

[143] WHO. Tuberculosis. Fact sheet No. 104, November 2010.

[144] M. D. Woolard and J. A. Frelinger. Outsmarting the host: bacteria modulating the immune response. *Immunologic Research*, 41(3):188–202, 2008.

[145] Z. Xiang, Y. Tian, and Y. He. PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biology*, 8(7):R150, July 2007.

[146] W. Xolalpa, A. J. Vallecillo, M. Lara, G. Mendoza-Hernandez, M. Comini, R. Spallek, M. Singh, and C. Espitia. Identification of novel bacterial plasminogen-binding proteins in the human pathogen *Mycobacterium tuberculosis*. *Proteomics*, 7:3332–3341, 2007.

[147] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928–942, 2004.

[148] R. J. Youle and A. Strasser. The BCL-2 protein family: opposing activities that mediate cell death. *Nature Reviews Molecular Cell Biology*, 9:47–59, January 2008.

[149] D. Young and T. Garbe. Lipoprotein antigens of *Mycobacterium tuberculosis*. *Research in Microbiology*, 142(1):55–65, 1991.

[150] N. Y. Yu, J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. C. Sahinalp, M. Ester, L. J. Foster, and F. S. L. Brinkman. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615, 2010.